La investigación sobre biblioteca digital. Pasado, presente y prospectiva

Georgina Araceli Torres Vargas
COORDINADORA

Publicación conmemorativa del X Aniversario del Instituto de Investigaciones Bibliotecológicas y de la Información: "A 40 años de investigación en Bibliotecología e Información en la UNAM"

Diseño de portada: Mario Ocampo Chávez

Primera edición: 18 de abril de 2023

D. R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
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

Esta edición y sus características son propiedad de la Universidad Nacional Autónoma de Mexico. Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titulr de los derechos patrimoniales.

Impreso y hecho en México

Contenido

Los datos y su estudio en la bibliotecología

Juan Voutssás-M.
Instituto de Investigaciones Bibliotecológicas y de la Información
Universidad Nacional Autónoma de México

Introducción

El estudio de los datos no es fenómeno nuevo; los datos siempre han sido la fuente primaria de la información, pero durante siglos el énfasis estuvo en esta última; los datos únicamente han sido considerados como una materia prima en su proceso. En los últimos años esta situación ha cambiado, y los datos han cobrado especial relevancia. Para definirlo de manera 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 lo mismo, los datos se colectan y se agrupan.

A lo largo de los últimos siglos, los resultados de estudios e investigaciones se presentaban como documentos "terminados": libros, revistas, tesis, etcétera, mientras que los datos utilizados para ellos generalmente se descartaban. Durante la segunda mitad del siglo xx, con el advenimiento y el desarrollo de las computadoras, se fue gestando el fenómeno del procesamiento de datos, el cual fue tomando, cada vez más, dimensiones inéditas. Nacieron así la gestión de datos, la ingeniería de datos, el análisis de datos, etcétera, todo lo cual fue creando además una "ciencia de los datos" para su estudio y desarrollo. Se le

atribuye a John Tukey[1] haber expresado, a principios de los sesenta, las primeras ideas que dieron origen a esta ciencia; durante esos años se creó también el concepto y la tecnología de la "base de datos", muy utilizada en nuestros días. Poco después, a principios de los setenta, Peter Naur publicó un texto que ya describe a la "Datalogía, la ciencia de los datos y de los procesos de datos y su lugar en la educación"; a lo largo de esa obra, el término "ciencia de los datos" ya fue utilizado ampliamente. El autor ofreció además una primigenia definición para esa "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".[2]

Obviamente este concepto ha evolucionado; hoy en día la "ciencia de los datos" (*data science*), comprende muchos más campos: desde las fuentes de datos, su selección y colecta, su gestión y preservación; la creación de "almacenes de datos", la aplicación de técnicas para la minería de datos; los repositorios de datos, la detección de tendencias en redes sociales, la interacción entre hombre y computadora, la visualización y análisis de datos, la evaluación de su calidad y de la información derivada de ellos, su uso ético, la gobernanza de datos, hasta el diseño de políticas y legislaciones al respecto. La ciencia de los datos moderna consiste en el estudio de 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.

Los repositorios de datos de investigación

Asimismo, como derivación del gran desarrollo de los datos procesados electrónicamente, a principios de este siglo se sumó otro fenómeno: la reconceptualización de la ciencia. Por siglos, ésta se basó en dos paradigmas fundamentales: la teoría y la experimentación. Como resultado del auge de las computadoras en la segunda mitad del siglo pasado, se integró un tercer paradigma: el modelado y la simulación con estos equipos. Pero eso no quedaría ahí: a comienzos de este siglo, Jim Gray afirmaba que había un cuarto paradigma para la ciencia contemporánea, el cual complementaba a los tres anteriores: los datos. La

1 John W. Tukey, "The Future of Data Analysis". *The Annals of Mathematical Statistics*, 33, num. 1 (1962), 1-67. doi:10.1214/aoms/1177704711

2 Peter Naur, *Concise Survey of Computer Methods* (Studentlitteratur: Lund, Sweden: 1974), [s. p.]. Citado por: Gil Press, "A Very Short History of Data Science". *Revista Forbes*, May 28 (2013). https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf.

ciencia se basaba tan profundamente en ellos que había que reconceptualizarla. Por lo mismo se requería toda una nueva generación de conceptos, herramientas, metodologías y expertos para tratarlos y, en especial, para estudiarlos. Sus teorías fueron recogidas por Hey *et al.*[3] en una primera antología sobre el tema, considerada la piedra angular de nueva visión de la ciencia basada en los datos. Además, Carlson[4] y Hey y Hey[5] señalaron desde ese año que las facetas emergentes de la ciencia —la e-ciencia o ciencia electrónica, la ciencia abierta, el manejo y re-uso de los datos científicos— habían establecido ya una nueva relación entre la ciencia y la biblioteca derivada precisamente de los datos.

A partir de entonces y cada vez más, los repositorios con datos provenientes de las investigaciones se han convertido en un gran insumo que requiere administrarse, distribuirse y preservarse de manera correcta, y para ello, obviamente, se requiere de la bibliotecología. Numerosas instituciones de investigación acudieron a sus bibliotecas para asesoría y para que ellas comenzaran a alojar esos conjuntos de datos, con lo cual se integraron así los repositorios de datos a sus colecciones usuales, y esto representó un reto inédito. Diversas organizaciones bibliotecarias comenzaron a estudiar estos nuevos retos, como la Asociación de Bibliotecas Universitarias y de Investigación (Association of College and Research Libraries: ACRL) de Estados Unidos, la cual es una subdivisión de la American Library Association (ALA),[6] así como la Liga de Bibliotecas Europeas de Investigación (Ligue des Bibliothèques Europeénnes de Recherche: LIBER).[7] Igualmente lo hizo la Federación Internacional de Asociaciones de Bibliotecarios y Bibliotecas (IFLA). En el último fascículo de 2016 y en el primero de 2017 de su revista,[8] compiló alrededor de 20 textos y re-

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

4 Scott Carlson, "Lost in a Sea of Science Data". *The Chronicle of Higher Education*. June 23 (2006) [s. p.]. https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

5 Tony Hey y Jessie Hey, "E-Science and its implications for the library community". *Library Hi Tech*, 24, núm. 4 (2006), 525-526. http://www.emeraldinsight.com/doi/pdfplus/10.1108/07378830610715383

6 Carol Tenopir, Ben Birch y Suzie Allard, *Academic Libraries and Research Data Services: Current Practices and Plans for the Future* (Association of College and Research Libraries (ACRL), 2012). http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf; Corol Tenopir, *et al*., "Research Data Services in Academic Libraries: Data Intensive Roles for the Future?" *Journal of eScience Librarianship*, 4, núm. 2 (2015). https://escholarship.umassmed.edu/jeslib/vol4/iss2/4/

7 Carol Tenopir *et al.*,"Research Data Services in European Academic Research Libraries". *LIBER Quarterly*, 27, núm. 1 (2017), 23-44. doi: http://doi.org/10.18352/lq.10180

8 *IFLA Journal*, 43, núm. 1 (2016); *IFLA Journal*, 42, núm. 4 (2017). https://www.ifla.org/publications/node/1691

flexiones al tema de los repositorios de datos, dividiéndolos en cuatro grandes temas: 1) las necesidades de los investigadores, 2) las habilidades requeridas de los bibliotecarios, 3) los posibles servicios a ofrecer y 4) la alfabetización en datos. A partir de esos estudios preliminares, la IFLA creó una iniciativa llamada Proyecto del Curador de Datos (Data Curator Project),[9] cuyo objetivo principal fue establecer con claridad las funciones y responsabilidades de los profesionales bibliotecarios que ya trabajaban en ello en diversos países, así como unificar la terminología utilizada para describir las nuevas funciones profesionales y prácticas emergentes. Witt y Horstmann[10] encontraron que las tareas primordiales requeridas a los bibliotecarios al respecto son: 1) ayudar a los investigadores a entender y resolver las necesidades a lo largo del ciclo de vida de los datos de las investigaciones; 2) asesorar en la construcción de planes de gestión de datos y metadatos; 3) diseñar soluciones de publicación y conservación de datos; 4) crear guías y tutoriales web para capacitar a investigadores y usuarios; 5) alojar y mantener repositorios en sus acervos.

Además, todos estos nuevos conceptos, proyectos y responsabilidades crearon una nueva especialidad en el campo de la información denominada Gestión de Datos de Investigación (Research Data Management: RDM). Whyte y Tedds la definieron así: "[…] la organización de los datos, desde su entrada en el ciclo de investigación hasta la difusión y el archivado de los resultados valiosos".[11]

De manera general, la RDM tiene que ver con todos los aspectos relativos a los datos emanados de la investigación: su ciclo de vida; colecta, depuración, coherencia y normalización; sus formatos, metadatos, los repositorios y servicios de consulta de datos, seguridad y anonimato de datos; su preservación, las habilidades y funciones requeridos para su gestión, alfabetización en datos para investigadores, e inclusive citación. Pinfield *et al.*[12] establecieron siete grandes campos de desarrollo, que denominaron "impulsores", para el estudio de la Gestión de Datos de Investigación: almacenamiento, seguridad, preservación, cumplimiento de políticas y leyes, calidad, difusión y compromiso.

Siendo una de las aplicaciones más conocidas, el manejo y estudio de los datos desde la bibliotecología no se limita a estos repositorios académicos.

9 IFLA, The Data Curator Project. https://www.ifla.org/library-theory-and-research/projects

10 Michael Witt y Wolfram Horstmann, "International approaches to research data services in libraries". *IFLA Journal*, 42, núm. 4 (2016), 251. doi: 10.1177/0340035216678726

11 Angus Whyte y Jonathan Tedds, *Making the case for research data management.* DCC Briefing Papers (Edinburgh: Digital Curation Centre, 2011), 27. https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm

12 Stephen Pinfield *et al*., "Research Data Management and Libraries: Relationships, Activities, Drivers and Influences". *PLOS ONE*, 9, núm. 12 (2014). https://doi.org/10.1371/journal.pone.0114734

Los datos se utilizan cada vez más en muy diversas facetas relacionadas con las bibliotecas y la industria de la información, pero, además, los datos se han ido sofisticando y multiplicando en proporciones inéditas (por lo que ya existen variantes de ellos), cada una de las cuales se ha convertido en una subdisciplina y una especialidad: los datos enlazados (*linked data*), los datos abiertos (*open data*), los datos masivos (*big data*), etcétera. Todos ellos tienen inmensas aplicaciones potenciales en las bibliotecas, y por tanto son materia de estudio desde la bibliotecología.

De cada una de estas subdisciplinas existen ya innumerables tratados y obras. No es intención de este texto hacer un análisis y compilación detallados, sino hacer un recuento básico para señalar su existencia y su importancia dentro del campo de la bibliotecología.

Datos enlazados

Básicamente, la teoría de los datos enlazados 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 personas, eventos, cosas, lugares, etcétera, estén conectados entre sí de manera estructurada, más poderosa será la red de datos, independientemente de que provengan de diversas fuentes y de que sus formatos no sean homogéneos. El propósito de esta teoría es potenciar el descubrimiento de conocimientos y la eficacia del análisis de datos.[13] Dada la importancia del tema en las bibliotecas, la IFLA[14] ya ha compendiado también el uso de los datos enlazados en ellas, y creó un grupo al efecto. Ávila estableció acerca del tema:

> […] la integración de los datos enlazados en los registros bibliográficos 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.[15]

13 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. Ahí, él menciona un estilo de publicación en la web con datos estructurados interrelacionados http://www.w3.org/DesignIssues/LinkedData.html.

14 *IFLA Journal*, 42, núm. 4 (2017). https://www.ifla.org/publications/node/1691

15 Eder Ávila, *Los datos enlazados y su uso en bibliotecas* (Ciudad de México: UNAM, Instituto de Investigaciones Bibliotecológicas y de Información, 2020), 80. https://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/56

Como ejemplos muy representativos de metadatos usando datos enlazados, la Biblioteca Británica y la Biblioteca del Congreso de Estados Unidos han comenzado a procesar los vínculos de datos entre sus cientos de respectivas colecciones —que comprenden muchos millones de ítems— para tratar de modelar las interrelaciones entre personas, eventos, lugares, etcétera, contenidos en sus acervos. Como ejemplos representativos del tema, se distinguen el Modelo de datos para libros de la Biblioteca Británica (British Library Data Model: Books),[16] y el Servicio de Datos Enlazados de la Biblioteca del Congreso de Estados Unidos (Library of Congress Linked Data Service).[17]

Datos abiertos

Se conoce como "datos abiertos" (*open data*) a un conjunto de iniciativas gestadas en años recientes bajo esta denominación común. Tienen como fin impulsar la creación, difusión y uso de repositorios de datos de todo tipo con acceso libre y abierto. Dichas iniciativas son una extensión de otros movimientos previos en favor de la apertura: *software* libre, Gobierno Abierto, revistas académicas abiertas, etcétera. Su importancia radica en que, además de los datos de investigación académica, otros tipos de datos abiertos se convierten en material de investigación relevante a nivel mundial, y permiten una comprensión más integral de los problemas y cuestiones globales relevantes, como educación, trabajo, salud, seguridad, o cambio climático. Son elemento indispensable dentro de los principios de Gobierno Abierto para la transparencia y rendición de cuentas, ya que empoderan a los ciudadanos y, por tanto, fortalecen la democracia participativa. Los datos abiertos tienden a optimizar los procesos y estructuras sociales que gobiernos y sociedad van construyendo, y se constituyen como apoyos esenciales para movimientos en favor de la igualdad racial, de género, etcétera. En suma, pueden ayudar a transformar la forma en que entendemos el mundo moderno y nos relacionamos con él.[18]

Ya existen algunos proyectos representativos operando al respecto, como los Datos Abiertos del Banco Mundial (World Bank Open Data), cuyo repositorio contiene en forma abierta más de 3 000 conjuntos de datos globales acerca de economía, demografía, desarrollo, etcétera. Con respecto al sector salud, se encuentra el Repositorio de Datos Abiertos de la Organización Mundial de la Salud, el cual compila información estadística sobre este tema proveniente de sus casi 200 miembros. Existe también el Portal de Datos Abiertos de la Unión

16 https://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf
17 https://id.loc.gov/
18 https://opendatacharter.net/principles-es/

Europea, con 12 000 conjuntos de datos provenientes de gobiernos, agencias, instituciones, etcétera, de esa región. Hay también proyectos al respecto de organizaciones de información, como DBPedia de Wikipedia, Registro de Datos Abiertos en Recursos de la plataforma AWS (Registry of Open Data on AWS Resources: RODA) de Amazon, el Explorador de Datos Abiertos de Google (Google Public Data Explorer), por mencionar algunos. Existen también 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 compilan y dan acceso a, por ejemplo, datos climáticos y meteorológicos de toda Norteamérica; y como ellos, los servicios sismológicos, vulcanológicos, censales, entre otros, a nivel regional o local en numerosos países.

Lo relevante de este tema desde el punto de vista de la bibliotecología es que el diseño, gestión y distribución de este tipo de datos abarcan un campo infinitamente mayor que sólo la gestión de datos de investigación. Aun si esos proyectos no se gestan o se insertan directamente en una biblioteca, requieren indefectiblemente de personal con conocimientos y experiencia en la gestión de datos. Buena parte de esos desarrollos ya involucran a personal proveniente de bibliotecas, pero evidentemente podrían ser más. El personal profesional bibliotecario, sin duda, tiene en esos proyectos de datos abiertos, más allá de los de investigación académica, grandes oportunidades de desarrollo profesional.

Datos masivos

Además de las anteriores, una vertiente muy especial de los datos son los datos masivos (o *big data*). Éstos no deben verse como una simple evolución vertical del concepto de los datos a lo largo de las décadas, sino la conjunción simultánea de múltiples fenómenos, teorías, necesidades, métodos y herramientas tecnológicas relacionados con los datos y la información que en un cierto momento se fueron convirtiendo en algo mucho más complejo. Como muchas otras tecnologías, los datos masivos fueron presentados y tratados por años como una panacea del manejo de datos e información y, por lo mismo, crearon en personas y organizaciones demasiadas expectativas; hoy en día, su tratamiento toma ya las dimensiones y perspectivas reales. Empero, los datos masivos representan ciertamente una herramienta válida para el análisis de información con miras a la toma de decisiones en las organizaciones, y obviamente entre ellas están las bibliotecas y archivos. Su conocimiento y manejo no pueden dejarse únicamente a cargo del personal de informática: es indispensable que bibliotecarios y archivistas también se adentren en ello, ya que

es un factor de valor agregado tanto para las organizaciones como para su personal dedicado a la gestión de información.

El fenómeno de los datos masivos se derivó del gran crecimiento de la red mundial y de las telecomunicaciones a partir de la década de los noventa, lo cual detonó un inusitado crecimiento de la información y los datos en su forma digital, especialmente impulsado por el auge de las redes sociales y de los dispositivos conectados a la red: el "internet de las cosas" y los *wearables*, o dispositivos conectados a la red. Millones de ítems en estas modalidades se sumaron a lo ya existente, con lo cual se multiplicó exponencialmente la cantidad de información acumulada. Aunado a esto, debe agregarse el gran negocio multimillonario creado en las dos últimas décadas alrededor de los datos, así como el incremento inusitado en la capacidad de almacenamiento de datos y su abaratamiento, todo lo cual contribuyó aún más en su crecimiento.

Básicamente, los datos masivos consisten en: 1) el tratamiento y análisis de conjuntos de datos tan grandes, variados, grandes, variados, complejos y dispares, 2) producidos a una velocidad vertiginosa y provenientes de muy diversas fuentes, 3) que hacen que los equipos, programas y procedimientos "tradicionales" de procesamiento de información (servidores, bases de datos, buscadores, etcétera), no sean suficientes, y 4) por tanto requieran de 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.

Los datos, y en especial los masivos, han sido utilizados cada vez más en años recientes en las bibliotecas y la industria de la información, en temas mucho más numerosos de los que aparentan a simple vista. Por lo general, los bibliotecarios y archivistas conocen los proyectos estandarte, por ser los más representativos al respecto, tales como el catálogo mundial Worldcat, operado por OCLC, que contenía, en 2019, 450 millones de registros catalográficos en casi 500 idiomas provenientes de cerca de 18 mil bibliotecas del mundo, y consigna también inventarios de 2 800 millones de obras en esas bibliotecas. Otro esfuerzo colectivo bibliotecario de datos masivos es la biblioteca digital HathiTrust: en 2020 este sitio consignaba ya más de 17 millones de ítems digitalizados. Se encuentra también en estos ejemplos el acervo denominado Archivos de Internet (*Internet Archives*), sitio sin fines de lucro que comenzó en los noventa para guardar páginas web con el fin de preservarlas, pero posteriormente se extendió a otros contenidos digitales o digitalizados; actualmente 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 1015 bytes de datos.[19]

Estos son los proyectos más conocidos de datos masivos en bibliotecas y archivos, pero no son los únicos; hoy en día este tipo de datos se usa en numerosos proyectos; por ejemplo, en el desarrollo de nuevas taxonomías de la información y esquemas de metadatos. Ello, debido a que es imposible pensar en explotar datos —de cualquier volumen— sin contar con adecuados metadatos; sin éstos, los conjuntos de datos, en especial los masivos, son entidades con poca o nula utilidad, y aún han sido poco desarrollados. Algunos estudios señalaban que en el año 2014 se agregaban metadatos de una manera sistematizada sólo al 3% de la inmensa cantidad de datos que se estaba produciendo en el mundo; los estudios actuales señalan un 5%. Un ejemplo interesante y actual de estos desarrollos son los modelos conceptuales subyacentes de las RDA, las directrices, elementos de datos, e instrucciones para crear metadatos de recursos bibliotecarios y del patrimonio cultural correctamente formados. 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, avalados por la IFLA y compatibles con el Modelo de Referencia de Bibliotecas (*Library Reference Model*). El punto central de esto consiste en que llevar de la teoría a la práctica a cada uno de esos modelos conceptuales implica el manejo de grandes cantidades de datos.

Los datos, y en especial los masivos, se usan también en los estudios métricos de la información documental, en todas sus especialidades: bibliometría, informetría, bibliotecometría, así como en otras asociadas: cienciometría, webmetría, altmetría, archivometría. Todas aplican modelos y métodos matemáticos y estadísticos a las actividades bibliotecaria, bibliográfica, archivística, las redes sociales, la investigación en ciencias y humanidades, etcétera. Son otro ejemplo de la minería de datos aplicada. Esto se usa para análisis de textos, análisis visual, etcétera. También para medir el acceso y uso de las revistas científicas como una medida más exacta de su impacto en lugar del tradicional análisis de citas. Se usan para retroalimentar la eventual selección de suscripciones a las revistas, y para ayudar a la toma de decisiones acerca de su renovación o cancelación. Halevi[20] compiló un resumen muy interesante de los tipos de uso de los datos masivos en bibliometría, y los categorizó en cinco: citas, referencias, palabras clave, uso, y análisis de textos completos.

19 https://archive.org/about/

20 Gali Halevi, *Bibliometric Big Data and its Uses* (2014). https://repositorio.unal.edu.co/bitstream/handle/unal/21558/biliometricsbigdata.pdf

Los datos masivos se utilizan también en la biblioteca en uno de los subcampos de la Inteligencia Artificial (IA), el denominado aprendizaje de máquina (*machine learning*) 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 realizar diversas acciones opcionales sin la intervención humana directa; estos sistemas específicos reciben datos que pueden interpretar, y de ellos extraer patrones significativos. Esta actividad es usada hoy en día no tan sólo en bibliotecas, sino también 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, OCR, etcétera. Permiten extraer suficiente información coherente de documentos para proporcionar elementos valiosos para las personas, como los catalogadores, o para sistemas, como los descubridores de biblioteca (*library discoverers*), y es una de las herramientas para diseño de futuros catálogos y recuperadores de información. Otra aplicación más de IA en bibliotecas son los denominados sistemas expertos, utilizados en este sector desde los años ochenta, tratando desde entonces indización basada en el conocimiento, procesamiento de lenguaje natural, catalogación, consulta y recuperación de información, *trend topic*s, etcétera.

Los anteriores no son todos los usos de los datos masivos en bibliotecas, pero sí los más representativos. Su relevancia creció a tal grado que la IFLA, en respuesta a las conclusiones de su "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, con el propósito de que las bibliotecas no sólo fueran simples espectadoras del fenómeno, sino que se convirtieran en parte proactiva del movimiento de los datos. El grupo se instauró formalmente durante el WLIC 2015 en Ciudad del Cabo, y desde entonces ha realizado una serie de estudios, eventos y documentos sobre el tema.[21] La afamada revista de bibliotecología *Library Journal* ha dedicado numerosos artículos al tema a lo largo de los últimos años, y 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.

21 Véase: IFLA Big Data Special Interest Group. https://www.ifla.org/big-data

Los datos como problema de investigación

A pesar de todo lo anterior, muchos bibliotecarios ven todavía con escepticismo el campo de los datos, en especial los masivos. En una alta proporción no saben cómo introducirse al tema, y muchos de ellos se preguntan si deben, y hasta dónde y cómo, adentrarse en la gestión y uso de los datos en todas sus variantes en beneficio de sus instituciones y de sus usuarios. Como se desprende de lo analizado previamente, la ciencia de los datos ya no es un tema de estudio sólo desde las ciencias de la computación, la estadística, etcétera, como lo fue originalmente. Hoy en día, es también un campo de estudio desde la bibliotecología y, por lo tanto, debe ser investigado y desarrollado desde este enfoque. Muchas universidades ya han comenzado este estudio, así como grandes organizaciones bibliotecarias como la IFLA, la ALA, otras en Europa, etcétera. El Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM (IIBI) también se ha adentrado en la investigación del tema desde hace ya algunos años.

Las principales preguntas de investigación acerca de la ciencia y práctica de los datos desde el enfoque de la bibliotecología giran alrededor de los siguientes puntos:

1) ¿Cuáles es el estado de la cuestión acerca de los datos y su uso en bibliotecas e industria de la información?
2) ¿Cuáles son los tipos y categorías de datos, sus variantes, y cuál es su teoría, principios, paradigmas, etcétera, en relación con la bibliotecología y los estudios de la información?
3) ¿Cuál es la utilidad y posibles aplicaciones de los datos en bibliotecas e industria de la información?
4) ¿Cómo pueden aprovechar y aplicar los datos los bibliotecarios de instituciones académicas, científicas, empresas, gobierno, bibliotecas públicas, etcétera?
5) En general, ¿qué deben saber los bibliotecarios acerca de datos y sus variantes?
6) En particular, ¿qué deben aprender los bibliotecarios acerca de las herramientas y metodologías informáticas utilizadas para el tratamiento y análisis de datos?

Acerca de todos estos temas existen ya innumerables proyectos, desarrollos, textos, conferencias, etcétera, a nivel global que tratan de comprender el desarrollo de la ciencia y práctica de los datos desde este enfoque bibliotecológico. Todo ello, con el fin de comprender el fenómeno, encontrar los beneficios

para los usuarios, y explicar los conocimientos, habilidades, aptitudes y herramientas que los bibliotecarios deben ir adquiriendo para aplicarlos en sus instituciones. Muchos de los autores se refieren ya al análisis de datos aplicado en bibliotecas más puntualmente como análisis bibliotecario (*library analysis*), para señalar con esta acepción específicamente a esa importante actividad relacionada a los datos en este ambiente y hacer evidente que ya es una especialidad dentro de las bibliotecas. Existe una gran cantidad de textos editados al respecto por la American Library Association (ALA), lo cual da cuenta de la importancia actual del tema.

En general, un aspecto muy importante de los estudios relacionados con eso consiste en que resaltan que la clave del éxito en este tipo de proyectos no reside en la adquisición y uso de los cientos de herramientas tecnológicas disponibles al respecto, sino en la calidad, capacidad, e iniciativa de los recursos humanos dedicados a ello. Dichos estudios han identificado en general un grupo de seis "roles" o funciones principales en esta ciencia y práctica 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 se establecieron otros tres roles adicionales relacionados estrechamente con la disciplina bibliotecológica: bibliotecario de datos, archivista de datos, y gestor/curador de datos. A este respecto, los profesionales de la información debían desarrollarse en cinco aspectos fundamentales:

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

Más allá de esta formación integral en el campo de los datos, y en función del tipo de biblioteca y/o organización en la que trabajen, los bibliotecarios deben ir adquiriendo y profundizando ese conocimiento para poder participar y colaborar en su entorno específico, para lo cual existen muy diversos campos de aplicación: en instituciones de investigación, en instituciones de enseñanza superior, en las empresas en general, en las empresas de Servicios de Información Bibliotecarios (LIS), en ciencias sociales y humanidades, en la organización y registro de la información y taxonomías, en la sistematización de métodos de recuperación documental, en la bibliometría, en la curaduría de datos, en la administración bibliotecaria, en la docencia y, por supuesto, en la investigación. 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 una 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.

Cada uno de los anteriores campos enunciados abre una amplia posibilidad de proyectos de datos en las bibliotecas y organizaciones afines; indudablemente, el campo de acción en ellas es sumamente amplio. Y en todas y cada una de esas eventuales aplicaciones se requieren profesionales con conocimientos, habilidades, actitudes, experiencias, y dominio de herramientas específicas. Obviamente es imposible que una biblioteca tenga especialistas en cada uno de esos campos, pero también debe quedar claro que las bibliotecas actuales no pueden carecer en lo absoluto de personal especializado en ello. Por tanto, la biblioteca puede y debe ir formando a 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 con 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.

Conclusiones

La gestión de datos en todas sus modalidades —simples, enlazados, abiertos, masivos, etcétera— ha dejado ya de ser un tema tecnológico emergente para convertirse en toda una realidad aplicada. Si bien siguen existiendo mitos y exageraciones al respecto, ya es indudable que 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 construido numerosos grupos, estudios y recomendaciones al respecto. Es un hecho que los campos de aplicación de los datos dentro de las bibliotecas son muy variados.

Derivado de su desarrollo e importancia dentro del campo de la información, es un tema que no puede ni debe ser menospreciado por las bibliotecas, ni quedar al margen en una zona de confort; deben abordarlo sin dilación y proactivamente. No es un tema que deba ser agregado casualmente ni un capricho que eventualmente puede ser adoptado por las bibliotecas como curiosidad técnica: representa a la vez un enorme reto y una gran oportunidad.

Ello significa nuevos esfuerzos y reacomodos, gastos y molestias. Por un lado, implica que su personal bibliotecario debe adquirir nuevos conocimientos, habilidades, experiencia y actitudes para su correcto manejo; pero, por otro lado, representa nuevas e inmensas oportunidades para reposicionar a la biblioteca dentro de las responsabilidades y quehaceres contemporáneos de su comunidad. Por lo mismo, requiere de estructura organizacional y personal calificado para realizar la tarea adecuadamente, como muchas de las otras tareas sustantivas de la biblioteca.

Como corolario a lo anterior, se reitera el hecho de que en los proyectos de datos la principal clave del éxito no está en las herramientas informáticas —por muy importantes que éstas sean—, sino en el personal calificado. Por ello, en la actualidad, es altamente recomendable que los bibliotecarios profesionales comiencen a buscar una cierta formación, adiestramiento y experiencia en la ciencia de los datos, en su gestión, su análisis, sus usos y aplicaciones, su curaduría, etcétera.

Por todo lo anterior, es importante investigar acerca de los datos desde el enfoque bibliotecológico y de los estudios de la información, para poder ampliar el conocimiento teórico, las aplicaciones prácticas, la docencia, la formación profesional y la educación continua de los profesionales de la información, en beneficio de todos los usuarios de ésta, en todos los ámbitos y niveles.

Bibliografía

Ávila, Eder. *Los datos enlazados y su uso en bibliotecas*. Ciudad de México: UNAM, Instituto de Investigaciones Bibliotecológicas y de Información, 2020. https://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/56

Carlson, Scott. "Lost in a Sea of Science Data". *The Chronicle of Higher Education.* June 23 (2006). https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

Halevi, Gali. *Bibliometric Big Data and its Uses.* 2014. https://repositorio.unal.edu.co/bitstream/handle/unal/21558/bibliometricsbigdata.pdf

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

Hey, Tony; Tansley, Stewart; Tolle, Kristin (Eds.). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond,

Wa.: Microsoft Research, 2009. https://digital.library.unt.edu/ark:/67531/metadc31516/m2/1/high_res_d/4th_paradigm_book_complete_lr.pdf

IFLA Journal, 43, núm. 1 (2016). https://www.ifla.org/publications/node/1691

IFLA Journal, 42, núm. 4 (2017). https://www.ifla.org/publications/node/1691

Naur, Peter. *Concise Survey of Computer Methods*. Studentlitteratur: Lund, Sweden: 1974, [s. p.]. Citado por: Gil Press, "A Very Short History of Data Science". *Revista Forbes*, May 28 (2013). https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf

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

Stanton, Jeffrey. "Data Science: What's in it for the New Librarian?". *Infospace. The Official Blog of the Syracuse University iSchool*. July 16 (2012). https://ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/

Tenopir, Carol; Birch, Ben; Allard, Suzie. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future.* Association of College and Research Libraries (ACRL), 2012. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

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

Tenopir, Carol *et al.* "Research Data Services in European Academic Research Libraries". *LIBER Quarterly*, 27, núm. 1 (2017), 23-44. DOI: http://doi.org/10.18352/lq.10180

Tukey, John W. "The Future of Data Analysis". *The Annals of Mathematical Statistics*, 33, num. 1 (1962), 1-67. doi:10.1214/aoms/1177704711

Voutssás-M., Juan. 2022. *Datos masivos en bibliotecas / Big Data in Libraries* [Edic. bilingüe]. Ciudad de México:

UNAM, Instituto de Investigaciones Bibliotecológicas y de Información, 2022.

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

Witt, Michael; Horstmann, Wolfram. "International approaches to research data services in libraries". *IFLA Journal*, 42, núm. 4 (2016), 251-252. doi: 10.1177/0340035216678726

La investigación sobre biblioteca digital. Pasado, presente y prospectiva. Instituto de Investigaciones Bibliotecológicas y de la Información / UNAM. La edición consta de 50 ejemplares. Coordinación editorial: Anabel Olivares Chávez. Revisión especializada, corrección de pruebas y formación editorial: LOGIEM, ANÁLISIS Y SOLUCIONES S. DE R.L. DE C.V. Fue impreso en papel cultural de 90 g en los talleres de Migal Impresiones Digitales S.A. de C.V., 3er Anillo de Circunvalación, No. 73, colonia Barrio Santa Bárbara, Alcaldía Iztapalapa, Ciudad de México, C.P. 09000. Se terminó de imprimir en abril de 2023.