

EL MANEJO DE DATOS

Aproximación desde los estudios
de la información

Georgina Araceli Torres Vargas



Z666.73
L56M3

El manejo de datos. Aproximación desde los estudios de la información / Coordinadora Georgina Araceli Torres Vargas. - México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2019.

viii, 116 pp. - Colección: TECNOLOGÍAS DE LA INFORMACIÓN.

ISBN: 978-607-30-2690-1

1. Datos vinculados. 2. Minería de datos. 3. Investigación bibliotecológica.

I. Torres Vargas, Georgina Araceli, coordinadora. II. Ser.

Diseño de portada: Natalia Cristel Gómez Cabral

Primera edición, 2020

D.R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Ciudad Universitaria, 04510, Ciudad de México

Impreso y hecho en México

ISBN: 978-607-30-2690-1

Publicación dictaminada

2020

Contenido

Presentación.....	7
GEORGINA ARACELI TORRES VARGAS	

MINERÍA DE TEXTO Y MINERÍA DE DATOS

Identificación de los temas de investigación en los documentos científicos del Colegio de Postgraduados.	11
ÁNGEL BRAVO VINAJA	
SANTIAGO DE JESÚS MÉNDEZ GALLEGOS	
JORGE PALACIO NUÑEZ	

Minería de texto aplicada a un diagnóstico de usuarios en Ciencia y Tecnología: aprendizajes para fortalecer la investigación bibliotecológica.	31
LOURDES FERIA BASURTO	

Minería de Datos, el caso de estudio de la Biblioteca Dr. Valentín Gómez Farías de la Facultad de Medicina de la UNAM.	43
DAVID FLORES MACÍAS	
GUADALUPE VANESA CAROLINA GUTIÉRREZ HERNÁNDEZ	

SISTEMATIZACIÓN DE DATOS Y SERVICIOS DE INFORMACIÓN

Research Data Management and Libraries: Opportunities and Challenges.....	59
KRYSZYNA K. MATUSIAK	

Integración de los principios de <i>linked data</i> en el registro bibliográfico.....	75
---------------------------------------------------------------------------------------	----

EDER ÁVILA BARRIENTOS

Plan para el Desarrollo de la Ciencia de Datos y Big Data (PDCDBD) en la UNAM para fines académicos y administrativos.....	95
----------------------------------------------------------------------------------------------------------------------------	----

JAVIER SALAZAR ARGONZA

Plan para el desarrollo de la Ciencia de Datos y Big Data (PDCDBD) en la UNAM con fines académicos y administrativos

JAVIER SALAZAR ARGONZA
Universidad Nacional Autónoma de México

I. ANTECEDENTES

1. En años recientes, la UNAM ha comenzado a incursionar en varias líneas de trabajo y proyectos institucionales de índole académica y administrativa que involucran el uso de las tecnologías de Ciencia de Datos y Big Data. Dichas líneas y proyectos:

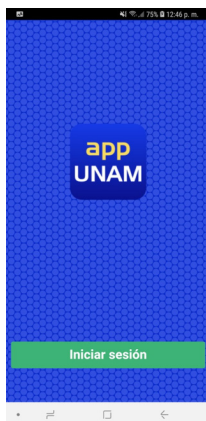
- Rebasan las capacidades de las herramientas disponibles en las áreas académicas y administrativas para su realización.
- Involucran el uso de *software* especializado (*frameworks*) y plataformas de cómputo de alto rendimiento (*clusters*), destinados hoy en día sólo a la investigación científica.
- Requieren de personal especializado (algo muy escaso).

2. Entre estas nuevas líneas de trabajo y proyectos institucionales, destacan:

A. La aplicación universal UNAM “AppUNAM” lo que:

- Permitirá recabar información estratégica de la comunidad universitaria, inclusive en tiempo real.
- Emplea dispositivos inteligentes.
- Analiza el *ClickStream*¹ con técnicas de Ciencia de Datos y Big Data, para abordar problemas antes irresolubles en relación con el aprendizaje y la eficiencia terminal de los estudiantes.

Figura 1. Pantalla de la AppUNAM.

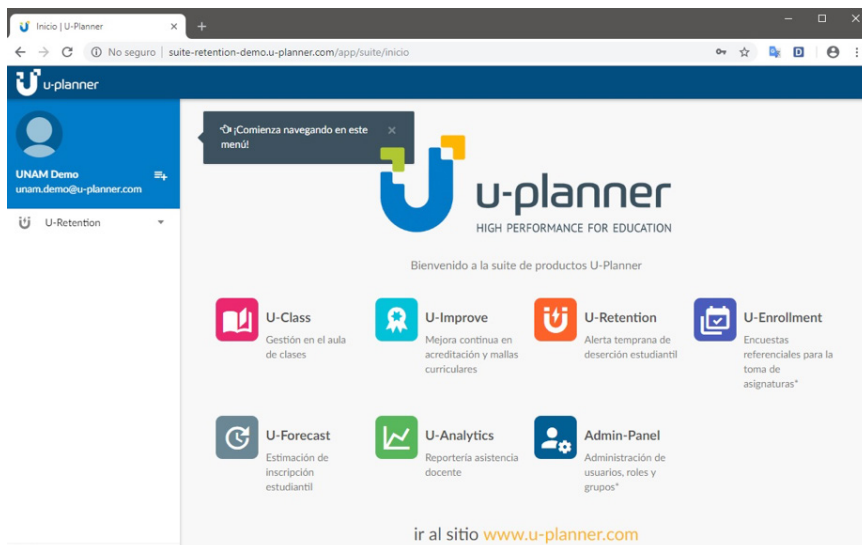


B. La adición de la UNAM al proyecto *Student Retention Workflow* de TANEQ. (Vía *U-planner*):

- U-planner permite cuantificar y combatir la deserción escolar.
- Emplea algoritmos de Inteligencia Artificial (*Machine Learning*).

1 ClickStream: Flujo de pulsaciones provenientes de los dispositivos inteligentes (Información).

Figura 2. Pantalla de la plataforma U-planner.



C. Programa de cuidado de la salud con IBM Watson, (Facultad de Medicina).

D. Proyectos de analítica del aprendizaje, (CUAED).

E. La nueva licenciatura en Ciencia de Datos en la UNAM.

F. Programas de capacitación y fomento de la cultura de Ciencia de Datos y Big Data para la comunidad universitaria, (Diversas dependencias).

3. Con una comunidad de 400 mil personas conformada por alumnos, profesores y trabajadores:

- La producción de datos masivos en la UNAM hoy en día ya es una realidad.
- Que requiere de las tecnologías de Ciencia de Datos y Big Data para su manejo y explotación.

Figura 3. Comunidad de la UNAM.
Fuente: <https://goo.gl/images/C79knF>.



4. La extracción de conocimiento a partir de los datos que se generan día con día en cada una de las áreas académicas y administrativas de la UNAM, resulta estratégica para:

- Mejorar la oferta educativa y la calidad de la enseñanza.
- Encontrar tendencias, desviaciones o irregularidades en la institución.
- Mejorar los procesos internos y los servicios.
- Diseñar nuevos servicios de aprendizaje personalizados.
- Conocer el sentimiento de la comunidad universitaria.
- Mejorar la seguridad de la información.
- Formar recursos humanos de excelencia en nuevas TIC, etcétera.

Figura 4. Extracción de Conocimiento.

Fuentes: <https://us.123rf.com/450wm/radiantskies/radiantskies1301/radiantskies130102072/17427648-abstract-word-cloud-for-knowledge-extraction-with-related-tags-and-terms.jpg?ver=6>
<https://sp.depositphotos.com/vector-images/extracci%C3%B3n-de-conocimiento.html>



5. Hasta hace algún tiempo las principales limitantes para utilizar las tecnologías de Ciencia de Datos y Big Data de forma regular en las áreas académicas y administrativas eran:

- Los costos y facilidades de acceso a las plataformas y recursos tecnológicos requeridos.
- La complejidad de las herramientas de *software*.
- La falta de personal especializado.
- La carencia de programas de capacitación.

Figura 5. Plataforma de Ciencia de Datos y Big Data.



6. Esta tendencia ha comenzado a cambiar hoy en día gracias a:

- La significativa reducción de costos en el *hardware* y *software* requeridos para hacer Ciencia de Datos y Big Data.
- El surgimiento de nuevas y mejores herramientas analíticas.
- La aparición de innovadores servicios de bajo costo en la nube.
- Mayor cultura informática.

Figura 6. Tendencias en la tecnología de Ciencia de Datos y Big Data.



7. Entre las principales estrategias que están comenzando a implementar las empresas e instituciones para utilizar Ciencia de Datos y Big Data, se tienen:

- La adquisición de plataformas y *clusters* dedicados al procesamiento y almacenamiento de datos.
- La adquisición de herramientas de analítica de auto-consumo (Power Bi, Tableau, Pentaho, etc.).
- La contratación de herramientas analíticas y de almacenamiento de datos en la nube (AWS, Google Cloud, Microsoft Azure, etc.).
- La contratación de servicios (DSaaS) “Ciencia de datos como servicio”.
- La capacitación y reclutamiento de personal (científicos de datos).

Figura 7. Herramientas tecnológicas actuales de Ciencia de Datos y Big Data.



II. ESTADO ACTUAL Y PROBLEMÁTICA

1. En los últimos veinticinco años se han instalado equipos, *clusters* de alto desempeño y supercomputadoras en diversas Facultades, Centros e Institutos de la UNAM (DSSI-DGTIC-UNAM 2018):

- Son equipos de propósito específico, excepto la supercomputadora.
- Permiten realizar trabajos de analítica y Big Data.
- Su uso está limitado a algunos cientos de proyectos de investigación científica al año.

Figura 8. Supercomputadora Miztli.

Fuente: <http://www.super.unam.mx/index.php/home/acerca-de?start=3>

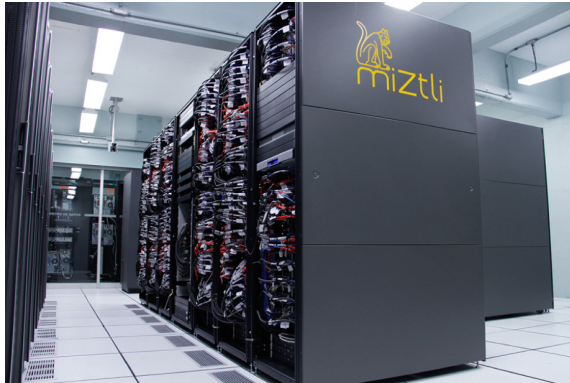


Figura 9. Cluster Instituto de Ciencias Nucleares.

Fuente: <http://www.nucleares.unam.mx/images/departamentos/altasenergias/FAE01.jpg>



2. En las áreas académicas y administrativas, se dispone de PCs, servidores y sistemas de información basados en un enfoque relacional y de inteligencia de negocios que no cuentan con las características técnicas para su uso en labores de Ciencia de datos ni de Big Data.

Figura 10. Equipo de cómputo del Instituto de Investigaciones Jurídicas.

Fuente: https://archivos.juridicas.unam.mx/www/site/generador/274Equipo_2164.JPG



Figura 11. Laboratorio de Cómputo de la Facultad de Ingeniería.

Fuente: https://hardwareviews.com/wp-content/uploads/2014/03/laboratorio-Nvidia-UNAM_a.jpg



3. En lo referente a la infraestructura disponible para la docencia en Ciencia de Datos y Big Data:

Manejo de datos...

- No se dispone de profesores con conocimientos en el tema.
- Se carece de aulas debidamente equipadas que faciliten la enseñanza de estas tecnologías.
- PCs o laptops con especificaciones avanzadas.
- Red de banda ancha.
- Acceso a *clusters* de alto rendimiento.
- *Software* especializado (Hadoop, Spark, Hive, Flume, Power BI, etc.).

Figura 12. Aula para Ciencia de Datos y Big Data.
Fuente: <http://www.gruposolutio.com/img/bigdata/dsl.png>



4. Los planes y programas de estudios disponibles en la UNAM para formar profesionales en el área de Ciencia de Datos y Big Data, actualmente no cubren por completo los temas de estudio que se requieren para este nuevo campo del quehacer humano.

Figura 13. Cursos aislados y programas de estudio que cubren parcialmente los temas de Ciencia de Datos y el Big Data.

Fuente: <http://www.unam.mx>

El mundo Big Data: Hadoop y Spark con Scala

Ciclo Integral "Inteligencia Artificial, Data Science, Deep Learning, TensorFlow"

Los temas Data Science (DS), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), TensorFlow, Keras, Spark y Python son hoy en día los temas de actividad en el mundo del desarrollo de aplicaciones productivas en tiempo real. En este ciclo integral se ofrecen los conceptos y herramientas de las nuevas tecnologías de Big Data y Data Science.

Fecha: 17 y 21 de abril 2018

Horario: Sábados de 8am a 3pm

No. de días: 2

Horas: 4

Lugar: Aula 101, edificio Vicerrectoral, Facultad de Ciencias, UNAM

Costo: \$12,000 pesos (más IVA)

Dirigido a: Directores, gerentes, analistas, consultores, ingenieros, programadores y estrategas de negocio interesados en incorporar e mejorar la práctica de la inteligencia de negocios en sus organizaciones, utilizando las mejores y más novedosas tecnologías existentes actualmente.

Objetivo: Proporcionar a los participantes los conocimientos necesarios que les permitan entender de una manera integral y objetiva, el nuevo enfoque del desarrollo y aplicaciones de la Inteligencia Artificial utilizando las mejores prácticas de Redes Neuronales Convolucionales como una extensión del Aprendizaje Automatizado (Machine Learning) en un ambiente de computo: Spark y con el lenguaje Python. Transmitir a los participantes el conocimiento de las mejores prácticas actuales del desarrollo BigData que utilizan las organizaciones y corporativas.

ESCUELA DE CIENCIA DE LOS DATOS

Segunda Escuela de Ciencia de los Datos. Una aproximación práctica

Ejes Temáticos:

- 1. Introducción a la ciencia de datos
- 2. Inteligencia artificial
- 3. Inteligencia de negocios
- 4. Inteligencia de datos

BIG DATA para BIBLIOTECARIOS ACADÉMICOS

14 de junio, 9am a 12 pm

INSTRUCCIONES PARA PARTICIPANTES

5. Existe una iniciativa para la creación de la licenciatura en Ciencia de Datos en la UNAM (México Nueva Era 2018). Se espera que sea capaz de cubrir las necesidades de los diferentes roles de personal que se requieren para trabajar la Ciencia de Datos y el Big Data. Participan:

- IIMAS.
- Centro Virtual de Computación.
- Ciencias.
- Ingeniería.
- Contaduría y Administración.
- Estudios Superiores Aragón.
- Institutos de Ingeniería II.
- Instituto de Ciencias Aplicadas y Tecnología.

6. La incorporación de la Ciencia de Datos y Big Data en las actividades cotidianas de las empresas e instituciones es ya una tendencia tecnológica mundial:

- En el 2017 un 40% de las empresas analizadas por Forrester Consulting, mostró que éstas ya disponen de

alguna estrategia enfocada al análisis masivo de datos (principalmente en sus áreas de mercadotecnia, desarrollo del producto y ventas).

- En el 2017, México se posicionó en segundo lugar en compras de soluciones de Big Data dentro de Latinoamérica, al adquirir el 26.7% del mercado, según la firma Frost & Sullivan (Olvera 2018). (El primer sitio lo obtuvo Brasil, con el 46.7% y el tercer lugar, fue Colombia, con el 7.9%).

7. La UNAM es líder en la formación y aprovisionamiento de recursos humanos altamente especializados, así como en el aprovechamiento y utilización de nuevas tecnologías:

Reconoce que la Ciencia de Datos y el Big Data constituyen hoy en día una de las herramientas más valiosas para elevar el nivel y proyección de la institución en los años por venir y propone impulsar su introducción y uso a través de un Plan de Desarrollo (PDCDBD).

III. SOBRE EL PLAN DE DESARROLLO PROPUESTO PDCDBD

- Es una iniciativa de la Dirección de Sistemas y Servicios Institucionales de la DGTIC.
- Busca atender los principales retos, a fin de que se desarrollen las tecnologías de Ciencia de Datos y Big Data en los ámbitos académicos y administrativos de la institución.
- Pretende reaprovechar los componentes útiles de la supercomputadora generación 5, que serán reubicados en el Centro de Datos de la UNAM.
- Está sustentado en el Plan para el Desarrollo del Supercómputo en la UNAM.
- Cumple con las directivas de:
 - El Plan de Desarrollo Institucional 2015-2019.
 - El Programa de Trabajo de Rectoría 2018.

- El Plan Maestro de Tecnologías de Información y Comunicación 2018.

IV. OBJETIVO

Proporcionarle a la comunidad universitaria:

- Recursos de cómputo para el desarrollo de proyectos de Ciencia de Datos y Big Data, dentro de un esquema eficiente, de calidad y pertinencia.
- Facilidades para extraer conocimiento de la información, sin importar lo compleja y voluminosa que ésta sea.
- Soporte en la toma de decisiones en todas las áreas del quehacer cotidiano de la universidad y del país.

V. METAS

- Abastecer, en la medida de lo posible, los requerimientos de la comunidad universitaria e incluso de otras instituciones y entidades nacionales y extranjeras en materia de Ciencia de Datos y Big Data.
- Iniciar la formación de especialistas que apoyen a la comunidad universitaria en el desarrollo de sus proyectos de Ciencia de Datos y Big Data, y que asesoren la implementación de estas tecnologías en otras instancias locales, regionales o nacionales.
- Implementar un modelo operativo y de negocios que genere recursos financieros para el crecimiento y actualización constante de los componentes necesarios para hacer Ciencia de Datos y Big Data en la UNAM.

VI. LÍNEAS ESTRATÉGICAS Y ALCANCES QUE SE CONTEMPLAN:

Tabla 1. Líneas estratégicas y alcances del Plan para el desarrollo de la Ciencia de Datos y Big Data en la UNAM para fines académicos y administrativos.

Línea.	Alcance.
✓ Infraestructura.	Disponer de los equipos y sistemas adecuados para atender las necesidades de Ciencia de Datos y Big Data de índole académica y administrativa de la UNAM.
✓ Capacitación.	Establecer los programas académicos de formación de especialistas y becarios.
✓ Servicios.	Brindar los nuevos servicios de Ciencia de Datos y Big Data a la comunidad universitaria.
✓ Desarrollo.	De la ciencia de datos y Big Data a nivel local, regional y nacional.
✓ Innovación.	Posicionar a la UNAM a la vanguardia de la Ciencia de Datos y el Big Data en México, Latinoamérica y el mundo.
✓ Marco normativo.	Que cubra un adecuado uso del hardware y software, manejo de información, garantice la actualización constante de los recursos, etc.

VII. ESCENARIOS DE SERVICIO POSIBLES (A, B, C) ²

Tabla 2. Escenarios de servicio posibles.

No.	Servicio.	A	B	C
1	Aprovisionamiento de infraestructura de hardware y software. (Vía el Centro de Datos de DGTIC). • Por medio de contenedores o máquinas virtuales, el lago de datos institucional y diversas herramientas de software colaborativo disponibles en la nube.	X	X	X
2	Mesa de ayuda vía ticket desde el Centro de Datos de DGTIC. • Sobre el aprovisionamiento de la infraestructura de hardware y software asignada y otros aspectos técnicos.	X	X	X
3	Soporte técnico.		Básico	X
4	Asesoría.		Básica	X
5	Consultoría para proyectos internos y externos.		Limitada	X
6	Cursos de capacitación.		X	X
7	Colaboración en proyectos internos y externos.			X

2 El escenario a utilizar dependerá de los recursos disponibles por DGTIC para la instrumentación de los servicios.

VIII. ETAPAS Y ACCIONES

i. Definición y diseño de la infraestructura requerida para brindar los nuevos servicios de Ciencia de Datos y Big Data:

- Determinar qué recursos del Centro de datos serán destinados para brindar los nuevos servicios.
- Identificar componentes de *software* y *hardware* adicionales.
- Planear el crecimiento del lago de datos institucional.
- Aspectos de seguridad de la infraestructura.
- Sistemas de apoyo (Tarificación, Mesa de ayuda, etc.).

Figura 14. Centro de Datos de DGTIC reacondicionado.



ii. Implementación de la infraestructura necesaria para brindar los nuevos servicios de Ciencia de Datos y Big Data

1. Realizar las adquisiciones, instalaciones y configuraciones necesarias en el Centro de Datos.
2. Designación, adecuación y amueblado de los espacios requeridos para brindar los nuevos servicios.
3. Adquisición del equipamiento y *software* del personal.
4. Reclutamiento y contratación del personal requerido.
5. Capacitación del personal.

6. Instrumentación del programa de becarios.
7. Desarrollo del marco normativo de servicios.
8. Creación de las redes de responsables y usuarios de Ciencia de Datos y Big Data.

iii. Consolidación de los servicios de Ciencia de Datos y Big Data para los fines académicos y administrativos que requiere la UNAM

1. Iniciar los servicios de Ciencia de Datos y Big Data para la comunidad.
2. Implementación del Lago de datos Académico-Administrativo de la UNAM con Acceso Abierto.
3. Iniciar las actividades académicas de Ciencia de Datos y Big Data.
4. Implementar el plan de negocios y comercialización de Servicios de Ciencia de Datos y Big Data.
5. Iniciar la sección de artículos y difusión de Ciencia de datos y Big Data en el portal de la UNAM.

iv. Innovación en Ciencia de Datos y Big Data

1. Crecimiento de la infraestructura destinada para Ciencia de Datos y Big Data en el data Center de DGTIC para fines académicos y administrativos.
2. Generación de un portal de auto aprovisionamiento de recursos de Ciencia de Datos y Big Data para usuarios internos y externos.
3. Creación de la Red Universitaria de Ciencia de Datos y Big Data abierta y distribuida.

IX. PERSONAL REQUERIDO PARA BRINDAR LOS SERVICIOS

Tabla 3. Personal requerido para brindar los nuevos servicios Ciencia de Datos y Big Data.

Plaza	Función
Responsable del Área y Líder de proyectos.	✓ Es responsable del área y control de los proyectos.
Arquitecto de sistemas.	✓ Establece la configuración de los recursos del Centro de Datos.
Administrador de plataforma.	✓ Administra los recursos de hardware y software en nube que son asignados a los usuarios finales.
Científico de datos. (Niveles Senior y Jr).	✓ Analiza datos, desarrolla algoritmos complejos e identifica oportunidades con técnicas estadísticas, algorítmicas de minería y visualización.
Científico de datos. (Nivel Citizen).	✓ Maneja herramientas de inteligencia de negocios "BI" que sirvan de interfaz con la con las herramientas de Ciencia de Datos y Big Data.
Ingeniero de datos.	✓ Carga información al ambiente de Big Data. ✓ Pone a disposición de los usuarios información, algoritmos y procesos de ciencia de datos.
Especialista en visualización.	✓ Convierte grandes volúmenes de datos en gráficos innovadores e instintivos.
Responsables de operación.	✓ Brindar servicios de apoyo en el Centro de Datos. (Tarificación, mesa de ayuda, etc.).
Servicios sociales y Becarios	✓ Desarrollo de algoritmos y programas específicos para manipulación de datos.

X. CURSOS DE FORMACIÓN PROPUESTOS PARA EL PERSONAL

1. Dirigidos al personal que conformará la nueva Área de Ciencia de Datos, con base en el rol y perfil que desempeñará en ésta.
2. Comprende veintinueve cursos distribuidos en ocho líneas de capacitación a lo largo de seis meses.

Tabla 4. Líneas de capacitación para el personal.

No.	Línea de capacitación	No. Cursos
1	Líder de proyecto de datos de Big Data y Ciencia de Datos.	9
2	Administrador de plataforma.	10
3	Ingeniero de datos.	16
4	Científico de datos.	13
5	Especialista en visualización.	4
6	Arquitecto de sistemas Big Data.	4
7	Responsables de operación.	4
8	Sensibilización a funcionarios.	2

3. Todos los cursos actualmente cuentan ya con su respectivo temario.
4. En un inicio el 100% de los cursos deberán ser adquiridos con proveedores externos.
 - La UNAM no cuenta con infraestructura de cómputo y personal capacitado para su realización.

XI. CONCLUSIONES

Es estratégico para la UNAM iniciar el aprovechamiento de la información que se genera día con día, en cada una de sus áreas académicas y administrativas a través de las tecnologías Ciencia de Datos y el Big Data.

El reaprovechamiento de los componentes de la supercomputadora Miztli, abre una excelente oportunidad para la UNAM de disponer de los recursos tecnológicos necesarios para comenzar a brindar nuevos servicios de Ciencia de Datos y Big Data a sus áreas académicas y administrativas.

La UNAM requiere de las tecnologías de Ciencia de Datos y de Big Data, para atender con eficiencia a su siempre creciente comunidad.

XII. BIBLIOGRAFÍA

- DSSI-DGTIC-UNAM. «Plan para el Desarrollo del Supercómputo en la UNAM 2018» (Documento interno en proceso de revisión para su publicación).
- UNAM. «Plan de Desarrollo Institucional 2015-2019». Acceso el 15 de Octubre de 2018. <http://www.rector.unam.mx/doctos/PDI-2015-2019.pdf>
- UNAM. «Programa de Trabajo de Rectoría 2018». Acceso el 15 de Octubre de 2018. <http://www.rector.unam.mx/doctos/Programa2018.pdf>
- UNAM. «Plan Maestro de Tecnologías de Información y Comunicación 2018». Acceso el 15 de Octubre de 2018. <https://www.red-tic.unam.mx/plan-maestroTIC.pdf>

Manejo de datos. Una aproximación desde los estudios de la información. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión especializada, Francisco Xavier González y Ortiz; revisión de pruebas, Valeria Guzmán González; formación editorial, Natalia Cristel Gómez Cabral. Instituto de Investigaciones Bibliotecológicas y de la Información / UNAM. Fue impreso en papel cultural de 90 gr. en los talleres de Grupo Fogra. Año de Juárez 223. Col. Granjas San Antonio. Alcaldía Iztapalapa. Ciudad de México. Se terminó de imprimir en febrero de 2020.