EL MANEJO DE DATOS

Aproximación desde los estudios de la información

Georgina Araceli Torres Vargas

La presente obra está bajo una licencia de:

https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es

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 lamisma licencia del original.

El manejo de datos. Aproximación desde los estudios de la información

COLECCIÓN
TECNOLOGÍAS DE LA INFORMACIÓN
Instituto de Investigaciones Bibliotecológicas y de la Información

El manejo de datos. Aproximación desde los estudios de la información

Coordinadora

Georgina Araceli Torres Vargas

Universidad Nacional Autónoma de México
2020

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

Ciudad Universitaria, 04510, Ciudad de México
Impreso y hecho en México
ISBN: 978-607-30-2690-1

Publicación dictaminada
2020

Contenido

Presentación

El término *dato* es impreciso; en algunos casos se refiere a la fuente primaria para respaldar una investigación, es decir a la evidencia para validar los resultados de investigación (Johnston, Lisa R. 2017, 2). Sin embargo existe una variedad de datos además de los que derivan de la investigación; el dato puede ser experimental, observacional, operacional, datos de un tercero, del sector público, datos de monitoreo, datos procesados o datos reutilizados (Austin, Claire C. 2016).

Tras la creciente proliferación de dispositivos móviles, transitan grandes cantidades de datos de diversa naturaleza a través de Internet. La coexistencia de esta heterogeneidad de datos es uno de los principales desafíos al momento de su manejo, por lo que surge una amplia diversidad de procesos para su análisis y sistematización, que va desde algoritmos genéticos, procesamiento del lenguaje, aprendizaje automático, redes neuronales, modelos predictivos, análisis de redes sociales, visualización de datos y minería de datos, por mencionar sólo algunos.

Desde los estudios de la información se ha vuelto necesario abordar cómo aprovechar las tecnologías y métodos que existen para efectuar el análisis de datos, con el fin de derivar servicios y productos de información acordes con los requerimientos que se tienen en el ámbito de la investigación, de la empresa, o de cualquier otro ámbito.

Frente a la amplitud de temas que circundan el estudio de los datos, la presente obra tiene por objetivo ofrecer algunas reflexiones en torno al tema del manejo de datos, que por lo general consta de la obtención de datos, su almacenamiento y su tratamiento.

En este sentido, se presentan tres trabajos relacionados con la minería de datos y de texto, que tienen como objetivo explorar el empleo de métodos para interpretar la información cualitativa, así como del análisis diacrónico de la producción científica.

De igual forma se presenta un capítulo relacionado con la gestión de datos de investigación, tema que surge como una nueva área de análisis y de práctica para los estudiosos de la información.

Otro aspecto es el referente a la adopción de los principios de *linked data* (datos enlazados), en la asignación de metadatos, para representar de forma granular los datos bibliográficos y su interrelación con otros datos en el entorno web.

El manejo de datos también requiere de conocimientos especializados, además de la infraestructura tecnológica. En el capítulo "Plan para el Desarrollo de la Ciencia de Datos y Big Data (PDCDBD) en la UNAM para fines académicos y administrativos" se muestran los antecedentes que motivan el desarrollo de este proyecto, así como la problemática que representa su implementación y el logro de sus alcances.

Como el lector podrá notar, un rasgo esencial de esta obra es que a lo largo de sus capítulos se reflexiona sobre las implicaciones sociales que trae consigo el manejo de los datos, así como los usos que y problemas que pueden derivarse. Las pautas de análisis que se dan para los datos, serán de utilidad para los estudiosos del tema y para quienes desean comenzar a adentrarse en la materia.

Austin, Claire C. "Key components of data publishing: using current best practices to develop a reference model for data publishing". En: *International Journal on Digital Libraries.* Junio 2016. Doi:10.1007/s00799-016-0178-2.

Johnston, Lisa R. "Introduction to data curation". En: *Curation research data. Volume One: practical strategies for your digital repository* / edited by Lisa R. Johnston, 2-24. Chicago, Illinois: Association of College and Research Libraries, 2017.

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

Ángel Bravo Vinaja
Santiago de Jesús Méndez Gallegos
Jorge Palacio Nuñez
Colegio de Postgraduados
Campus San Luis Potosí

INTRODUCCIÓN

El Colegio de Postgraduados (ColPos) es un Centro Público de Investigación dependiente de la Secretaría de Agricultura y Recursos Hidráulicos Pesca y Alimentación (SAGARPA), creado por Decreto Presidencial en 1959 (Colegio de Postgraduados 2014). En el 2001 el ColPos se constituyó en un Centro Público de Investigación, lo que le permitió autonomía y mayor independencia presupuestal, que cuando estaba bajo control de la SAGARPA y la Secretaría de Hacienda y Crédito Público (Colegio de Postgraduados 2016). Fue concebido como una institución pública estratégica para el desarrollo social del sector agropecuario y forestal de México, a través de la formación de recursos humanos de alto nivel, para generar información científica que contribuya al desarrollo y fortalecimiento de instituciones del sector (González

Cossío 2010). Su misión es "generar, difundir y aplicar conocimiento para el manejo sustentable de los recursos naturales, la producción de alimentos nutritivos e inocuos, y el mejoramiento de la calidad de vida de la sociedad" (Colegio de Postgraduados 2016). Esta institución imparte dieciséis programas de maestría y doctorado en ciencias en sus siete Campus, ubicados en los estados de: México, Puebla, San Luis Potosí, Tabasco, Veracruz (dos campus) y Campeche, los cuales son reconocidos por el Programa Nacional de Posgrados de Calidad (PNPC) del Consejo Nacional de Ciencia y Tecnología (Conacyt) (Colegio de Postgraduados 2017). En 2017 contaba con 444 profesores (de 616 plazas académicas) con grado de doctor en ciencias, formados en universidades de todo el mundo, de los cuales 56% pertenecían en ese año al Sistema Nacional de Investigadores (Colegio de Postgraduados 2016).

Varios de sus investigadores han sido reconocidos con el otorgamiento de premios internacionales, nacionales y estatales de ciencias y artes, en las áreas de tecnología y diseño y en ciencias naturales y exactas; premios de ciencia y tecnología de los alimentos y premios Banamex, entre otros. Además, es la institución de ciencias agrícolas mexicana que cuenta con el mayor número de investigadores nacionales Nivel III en el Sistema Nacional de Investigadores (SNI) (Larqué-Saavedra 2014). Pero la influencia del ColPos no se restringe a eso, ya que es pionero en la generación de conceptos y escuelas del pensamiento en las ciencias y tecnologías agrícolas, y por haber realizado aportaciones importantes para el desarrollo agrícola entre las que destacan: resaltar la importancia de la biodiversidad en México; establecer bancos de germoplasma, e implementar estudios fundamentales de los sistemas agrícolas y de los tipos de vegetación de México. Adicionalmente los investigadores han resaltado la importancia que representan los campesinos en la domesticación y conservación de las especies, la elaboración de mapas de suelos agrícolas y su conservación, así como el establecimiento de biofábricas de hongos comestibles y agentes de control biológico (Larqué-Saavedra 2014).

Actualmente, el ColPos cuenta con 49 líneas de investigación llamadas "Líneas de Generación y/o Aplicación del Conocimiento

(LGAC-CP)", donde confluyen la especialización de las investigaciones de los profesores-investigadores que conforman el Núcleo Académico Básico (NAB) de cada programa de postgrado, que son quienes definen la naturaleza de los programas de postgrado, además, de fundamentar los proyectos de investigación de los estudiantes y facilitar de esta forma la operación de la investigación (Colegio de Postgraduados 2018).

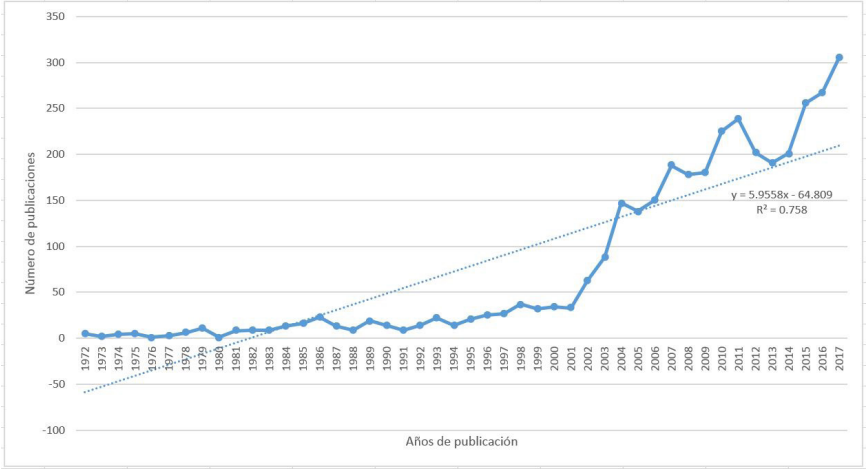
Los resultados de la actividad científica del ColPos, en sus primeros años de vida, no fue publicada en revistas internacionales indizadas en bases de datos analizadoras de la producción científica tales como el SCIE, el SSCI, contenidas en el Web de la Ciencia (WOS) y Scopus. En las bases de datos SCIE y SSCI, la primera contribución apareció hasta 1972; a partir de este momento, la publicación de contribuciones científicas hasta 1989 fue de 157 (4.54% del total publicado hasta 2017); es decir, 8.55 documentos por año. De 1990 a 2004 se publicaron 580 documentos (16.78% de lo publicado hasta 2017), esto es 38.66 por año. La mayoría de los documentos derivados de la investigación realizada en el ColPos se publicaron en revistas mexicanas, algunas de las cuales ahora aparecen listadas en el "Sistema de Clasificación de Revistas Mexicanas de Ciencia y Tecnología", en publicaciones seriadas del propio ColPos como "Comunicaciones en Estadística y Cómputo", "Cuadernos de Desarrollo Rural", "Comunicaciones en Socioeconomía, Estadística e Informática" y los primeros años de la revista "Agrociencia". Es en los últimos trece años que la actividad científica del ColPos se ve reflejada en las revistas de corriente principal, llamadas así por Salager-Meyer (2015) y a las revistas indizadas en las bases de datos SCIE y SSCI, ya que de 2005 a 2017 se indizaron allí 2 720 documentos del ColPos, que corresponden a 209.23 documentos por año.

En los últimos años los artículos indizados en las bases de datos SCIE y SSCI corresponden aproximadamente al 50% de la producción anual del ColPos, como lo muestran los datos estadísticos al respecto. En el año 2016 se indizaron 267 (48.72%) documentos en las bases de datos SCIE y SSCI, de 548 publicados en revistas con comité editorial reportados en el Sistema Integral de Informa-

ción Académica (SIIA) de esta institución. En 2017, la proporción subió a 51.26% (305 de 595); por lo tanto, se puede afirmar que las temáticas de investigación de la producción científica del ColPos de los últimos años que se analizan mediante minería de textos, corresponden en esta investigación al 50% de la producción total de la institución.

La producción científica del ColPos indizada en las bases de datos SCIE y SSCI, tuvo un crecimiento bajo de 1972 a 2000, pero a partir de 2001 comenzaron a indizarse un mayor número de publicaciones en las bases de datos mencionadas, presentando una tendencia creciente cada año, exceptuando un periodo entre 2011 a 2013, pero a partir de 2014 la indización de documentos volvió a crecer *Figura 1.*

Figura 1. Crecimiento de la producción científica del Colegio de Postgraduados en revistas de corriente principal de 1972 a 2017.

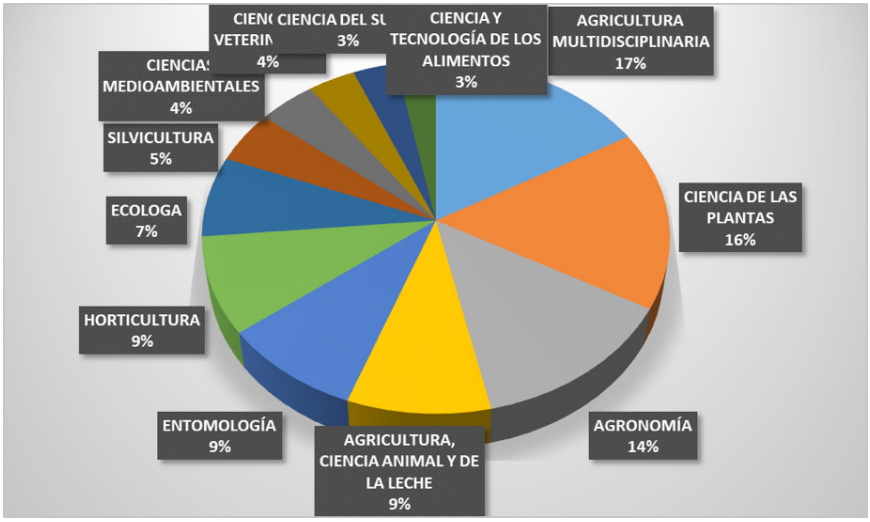


Los tipos de documentos indizados en las bases de datos SCIE y SSCI, en su mayoría, son artículos científicos (3 115, 90.17%), 199 resúmenes de congresos (5.76%), 42 editoriales (1.2%), 39 artículos *in extenso* (1.13%), 38 reseñas (1%), y el resto fueron noticias (32),

notas científicas (17), correcciones (5), biografías (3), reseñas de libros (3), cartas al editor (3), y un capítulo de libro. Vale la pena destacar la baja aportación del cuerpo académico del ColPos en la participación de libros indizados en estas bases de datos, ya que indizan principalmente revistas científicas.

En cuanto a las categorías temáticas del Web de la Ciencia, de las revistas en las que se indizaron las publicaciones del ColPos, 74% de los documentos corresponden a seis grandes temas de agricultura multidisciplinaria (17%), ciencias de las plantas (16%), agronomía (14%), ciencia animal y de la leche (9%), entomología (9%) y horticultura (9%), el resto (26%), estuvo distribuido en otras categorías tales como: ecología, silvicultura, ciencias medioambientales, ciencias veterinarias, ciencia del suelo, y ciencia y tecnología de los alimentos *Figura 2.*

Figura 2. Categorías temáticas de las revistas indizadas en el Web de la Ciencia de los documentos publicados por el Colegio de Postgraduados.



La minería de textos es el proceso de extracción de patrones o información interesante a partir de documentos de texto no estructurados (Tan 1999). En tanto que Feldman y Sanger (Feldman y Sanger 2006) lo definen como un intensivo proceso de conocimiento en el que un usuario interactúa con una colección de documentos mediante el uso de un conjunto de herramientas de análisis; mencionan además, que al igual que la minería de datos, la minería de textos busca extraer información útil de las fuentes de datos, sin embargo, en el caso de la minería de textos, las fuentes de datos son colecciones de documentos donde es posible encontrar patrones interesantes en los datos textuales no estructurados. Las aplicaciones de la minería de textos para encontrar patrones interesantes se dan principalmente, según Feldman y Sanger (2006) en áreas como la inteligencia de negocios o empresarial, el análisis de patentes, y la investigación en ciencias de la vida.

VOSviewer es un programa informático para construir y visualizar redes bibliométricas (CSTS 2018). Entre las múltiples tareas que éste puede realizar, se encuentra la minería de textos, la cual se puede realizar usando los títulos y resúmenes de los documentos. También ha sido utilizado como herramienta bibliométrica en diferentes documentos técnicos y de aplicación. Entre los documentos técnicos destacan trabajos de los creadores del programa de cómputo Ness Jan Van Eck y Ludo Waltman: "Text mining and visualization using VOSviewer" (Eck y Waltman 2007), y "VOS: A New Method for Visualizing Similarities Between Objects" (Eck y Waltman 2011). Respecto a documentos donde se aplica el análisis de textos usando Vosviewer, destaca el trabajo de Gobster (Gobster 2014) "(Text) Mining the LANDscape: Themes and trends over 40 years of Landscape and Urban Planning".

A partir de las facilidades que proporciona VOSviewer para realizar trabajos de minería de textos usando los registros bibliográficos de diferentes bases de datos como el Web de la Ciencia y Scopus, se están realizado trabajos usando la aplicación para identificar las temáticas y tendencias de investigación, como es el caso de este trabajo, que tiene como objetivo identificar y describir

las temáticas de investigación en los documentos publicados por el personal académico del ColPos en revistas de corriente principal, lo que servirá a los tomadores de decisiones del ColPos para afianzar o reorientar la investigación científica en la institución.

METODOLOGÍA

La búsqueda de la producción científica del Colegio de Postgraduados en revistas de corriente principal se efectuó en las bases de datos SCIE y SSCI del Web de la Ciencia de la empresa Clarivate Analytics mediante la ecuación de búsqueda mostrada en la *Figura 3,* limitando los resultados desde la publicación del primer documento en 1972 hasta el año 2017.

Figura 3. Ecuación de búsqueda de la producción científica del Colegio de Postgraduados en el Science Citation index Expended y el Social Sciences Citation index.

((AD=(COLEGIO POSTGRAD OR COLEGIO POSGRAD OR COLEGIO POSTGRADUADOS OR IREGEP OR COLEGIO POSTGRAD CIENCIAS AGR OR IRENAT OR COLEGIO POSTGRAD MONTECILLO OR COLEGIO POSGRADUADOS OR INST FITOSANIDAD OR INST RECURSOS GENET & PROD OR COLEGIO POSTGRAD CIENCIAS AGRICOLAS OR COL POSTGRAD OR COLEGIO POSTGRAD CHAPINGO OR COLEGIO POSTGRAD MICROBIOL EDAFOL IRENAT OR IFIT OR ISEI OR COLEGIO POSGRADUADOS MONTECILLOS OR COLEGIO POST GRAD OR COLEGIO POSTGRAD CARRETERA MEXICO TEXCOCO OR COLEGIO POSTGRAD CIENCIAS AGR MONTECILLO OR COLEGIO POSTGRAD CIENCIAS AGROCOLAS OR COLEGIO POSTGRAD CIENCIAS CIENCIAS AGR OR COLEGIO POSTGRAD CONACYT OR COLEGIO POSTGRAD EDO MEXICO OR COLEGIO POSTGRAD H CARDENAS OR COLEGIO POSTGRADOS OR COLEGIO POSTGRADUADOS CARRETERA MEXICO TEXCOCO OR COLEGIO POSTGRADUATOS OR COLEGIO POSTRAD OR CTR GANADERIA COLEGIO POSTGRAD OR "COLEGIO POSTGRAD, PROGRAMA GANADERIA" OR "PROGRAMA GANADERIA, KM 36.5" OR INST RECURSOS GENET PRODIOL OR INST RECURSOS NAT COLEGIO POSTGRAD OR INST SOCIOECON ESTADIST & INFORMAT OR IRENAT COLEGIO POSTGRAD OR IRENET OR IRGP OR CTR DOCUMENTAC & BIBLIOTECA, MONTECILLO OR "CAMPUS SAN LUIS POTOSI COLEGIO POSTGRAD" OR "CAMPUS SAN LUIS POTOSI,ITURBBIDE") AND CU=MEXICO) OR (OG=(Colegio de Postgraduados - Mexico))) NOT (AD=("UDG, COLEGIO POSTGRAD" OR "CELULOSA & PAPEL, COLEGIO POSTGRAD"))

Los registros bibliográficos de los documentos resultantes en idioma inglés se descargaron en una carpeta con el registro completo en texto sin formato; dichos registros se cargaron a VOSViewer

indicando el tipo de formato de acuerdo con la base de datos de procedencia. Los registros se separaron en tres periodos de años (1972-1989; 1990-2004; y 2005-2017). Para cada periodo se realizó la selección de los parámetros que solicita VOSviewer para realizar el análisis de minería de textos, los cuales dependen del número y tamaño de los archivos que se someten a análisis, y de la profundidad que se desee analizar y visualizar. Es decir, el número mínimo de veces que se repiten las palabras o frases identificadas por VOSViewer en el conjunto de registros; para el periodo 1972-1989, el número mínimo de repetición de las palabras o frases fue de dos; para el periodo 1990-2004, el número mínimo fue cinco, y para el periodo 2005-2017 fue de 10. Dentro del título o resumen de cada registro solo se tomó en cuenta una sola vez cada palabra. Para la normalización de los registros se usó un archivo en texto creado con las palabras o frases a normalizar y a excluir. El método de normalización de las palabras fue la "fortaleza de la asociación", opción por asignación usada en VOSviewer. Mediante la opción de visualización "overlay visualization" se identificaron los temas más actuales en el rango de años 2007-2015.

RESULTADOS

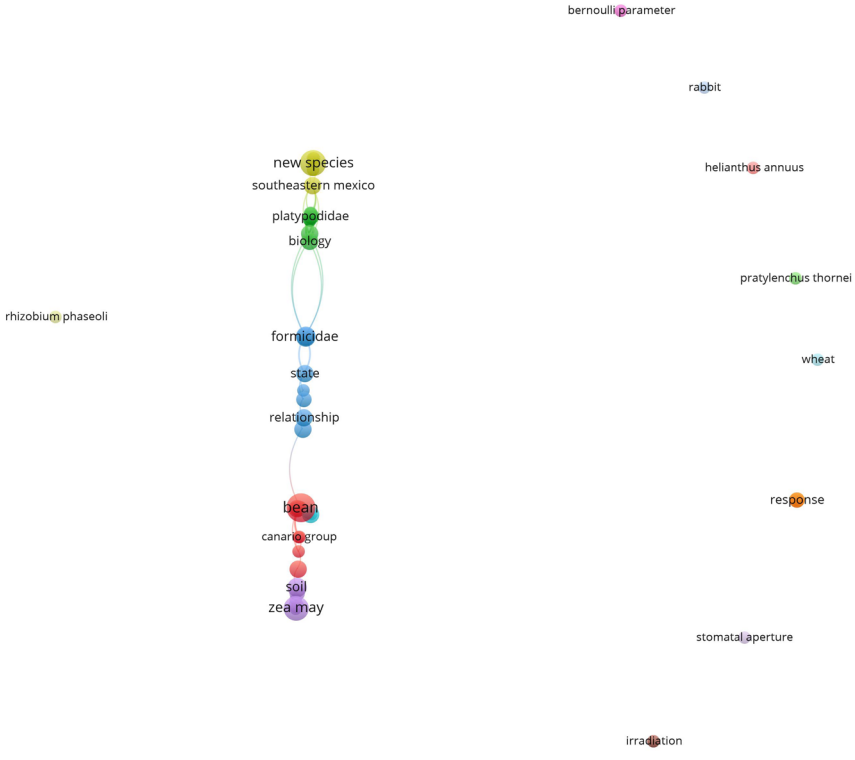
Se identificaron y descargaron 3 457 registros bibliográficos de las bases de datos SCIE y SSCI, y se separaron en periodos similares. De 1972 a 1989 se encontraron 157, que corresponden a 4.54%; del periodo 1990 a 2004 se encontraron 580 registros, que corresponden a 16.78%, y de los años 2005 a 2017 se encontraron 2 720 registros bibliográficos, que corresponden a 78.68% del total.

PERIODO 1972-1989

Se realizó en VOSviewer la minería de textos de los registros bibliográficos del periodo 1972-1989 obtenidos del título y resumen de dichos registros; sólo se tomó en cuenta una palabra o frase

por registro, lo que dio como resultado 537 ítems. El número mínimo de ocurrencias de las palabras o frases en el total de registros fueron dos; lo anterior dio como resultado 58 palabras o frases y se creó el mapa de red temático mostrado en la *Figura 4*.

Figura 4. Mapa de red de las temáticas obtenidas mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 1972-1989.



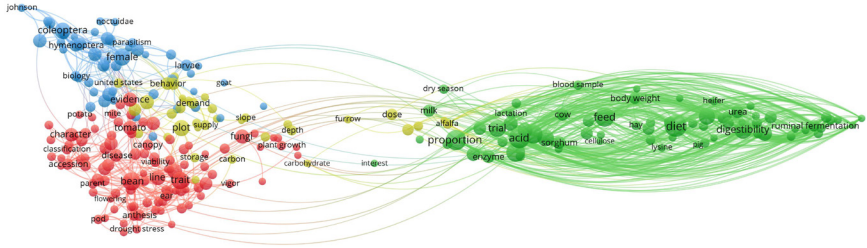
Se identificaron quince grupos temáticos, de los cuales sólo seis están interrelacionados: maíz y suelos, frijol muy relacionado con la acumulación de la fitohormona ácido abscísico, rendimiento, gramíneas y hormigas *(Formicidae),* biología de los escarabajos

ambrosiales, y el estudio de nuevas especies como las Lauráceas, nemátodos, y la Drosophila mexicana; también se identificaron nueve temas de investigación que no tienen relación entre ellos: la bacteria *Rhizobium phaseoli,* parámetros de Bernoulli, conejos, girasoles, el nemátodo *Pratylenchus thornei,* trigo, apertura estomatal, irradiación y respuesta *Figura 4.*

PERIODO 1990-2004

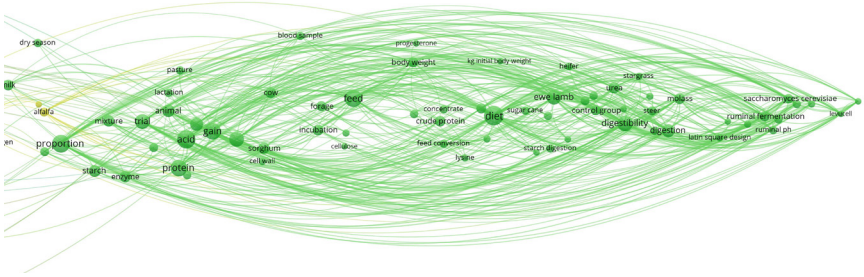
Se realizó minería de textos de los registros bibliográficos del periodo 1990-2004 obtenidos del título y el resumen de dichos registros de los cuales sólo se tomó en cuenta una palabra o frase por registro, lo que dio como resultado 12 472 ítems; el número mínimo de ocurrencias de las palabras o frases en el total de registros fue cinco. Lo anterior dio como resultado 429 palabras o frases, de las cuales se seleccionaron el 60% de los términos más relevantes, lo que resultó en 257 ítems o palabras para realizar el mapa de red temático mostrado en la *Figura 5.* En dicho mapa se identifican cuatro grupos temáticos con dos secciones claramente diferenciadas, por un lado, está un grupo que trata temas de ganadería, y por el otro, tres grupos o *clusters* con temas como producción de cosechas, suelos, y enfermedades y plagas de las plantas.

Figura 5. Mapa de red de las temáticas obtenidas mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 1990-2004.



En el grupo de ganadería, se identificaron temas de investigación como: materia seca, dieta, alimentación, digestión, digestibilidad, fermentación, fermentación ruminal, y forrajes como alfalfa, sorgo, paja de maíz, soya y pastos; otros temas son proteínas, almidón, leche, y levaduras, principalmente *Saccharomyces cerevisiae* y levucell. En cuanto a la investigación sobre animales criados para alimentación humana se destaca la investigación sobre ovinos, bovinos y cerdos. Otros temas destacados son: melaza, suplementación, progesterona, enzimas, lisina, microorganismos y bacterias *Figura 6.*

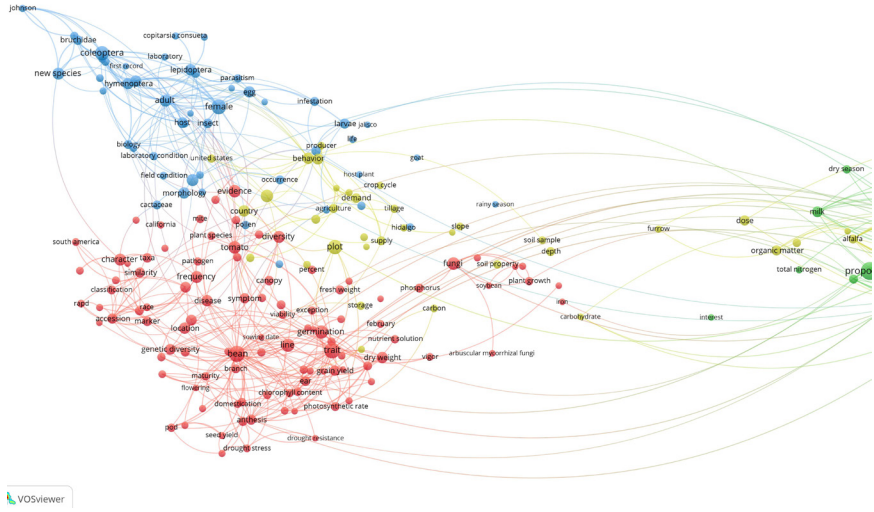
Figura 6. Mapa de red de las temáticas sobre ganadería, obtenidas mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 1990-2004.



En el bloque de tres grupos o *clusters,* se destaca la investigación sobre floración, domesticación, deficiencia y estrés a la sequía en frijol; y la condición y el cultivo de tomate en invernadero. También destacan las investigaciones sobre soya, papa, diversidad genética, germinación de semillas, hongos, hongos micorrícicos y fotosíntesis. Aparecen también en este *cluster* las investigaciones sobre enfermedades de las plantas, ácaros, nemátodos como *Naccobus aberrans,* diversas plagas como *Copitarsia consueta,* insectos como coleópteros (principalmente Brúquidos), himenópteros y lepidópteros. La morfología de cactáceas también es un tema destacado. Otros temas importantes son el muestreo y las propiedades de suelos, materia orgánica, labranza convencional, irrigación,

mercadeo de productos agrícolas, y elementos de importancia para la nutrición vegetal como el carbono, hierro y fósforo *(Figura 7).*

Figura 7. Mapa de red de las temáticas sobre producción de cosechas, suelos y plagas y enfermedades, obtenidas mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 1990-2004.

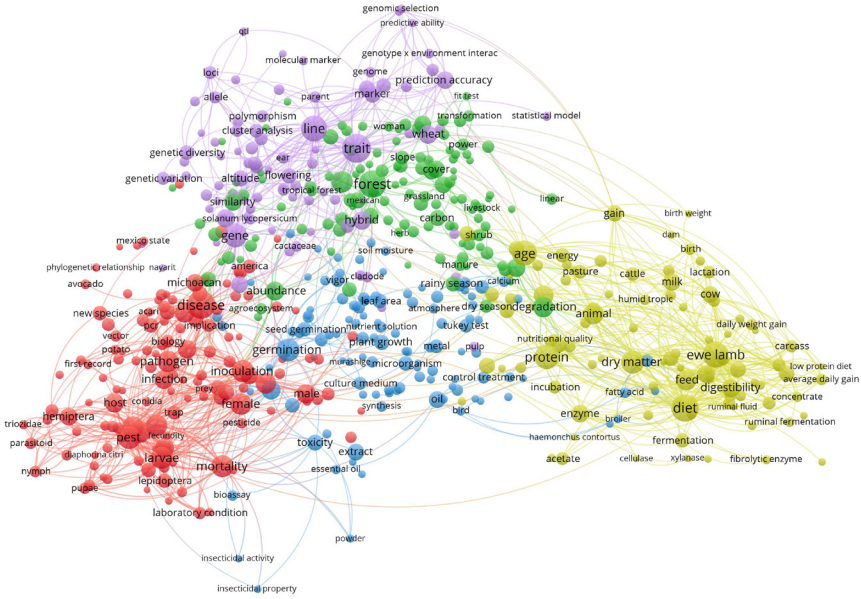


PERIODO 2005-2017

Del proceso de minería de textos aplicada al título y resumen de los documentos, indizados de 2005 a 2017, se obtuvieron 50 934 palabras o ítems, seleccionando las que tenían al menos 10 ocurrencias; resultaron 948 ítems, de las cuales se seleccionaron 60% de las que tienen mayor significancia. Lo anterior dio como resultado un mapa con 569 palabras o frases, con las que se formaron cinco *clusters* o grupos temáticos donde se aprecia un mapa con forma de triple hélice, donde el aspa izquierda contiene los temas relacionados con "producción animal" (color verde olivo), en

el aspa derecha (color rojo) se encuentra el grupo que trata sobre enfermedades y plagas de las plantas, en el aspa superior (color morado) se encuentra el *cluster* que trata sobre genética vegetal, y en medio de las hélices se encuentra dos grupos o *clusters* que tratan sobre suelos y bosques (color verde) y ciencias de las plantas (color azul) *(Figura 8).*

Figura 8. Mapa de red de las temáticas obtenidas mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 2005-2017.



En el grupo sobre producción animal (color verde olivo), se destaca la investigación sobre dietas, alimentación, suplementación, digestibilidad y desempeño del crecimiento en ovinos; dieta, suplementación y producción de leche en ganado vacuno; dieta, desempeño del crecimiento, y tamaño de la canal en cerdos, y la investigación en caprinos. De manera general, se destaca en este *cluster* la inves-

tigación sobre dieta, ingestión de proteína, digestión, materia seca, la fermentación, degradación ruminal, ganancia de peso, la percepción de la investigación en pastizales y sorgo, y el uso de enzimas para la fermentación de la materia seca *(Figura 8).*

En el *cluster* sobre plagas y enfermedades de las plantas (color rojo), se destaca la investigación insectos y otros organismos fitopatógenos, como ácaros y virus que afectan la producción de cultivos para la alimentación humana y animal; entre las plagas más importantes se destacan: el picudo del agave *(Scyphophorus acupunctatus),* el psílido asiático de los cítricos *(Diaphorina citri)* y el psílido de la papa y tomate *(Bactericera cockerelli);* se destacan también en este grupo los temas, sobre control biológico y control tradicional, así como el uso de los hongos entomopatógenos *Metarhizium anisopliae* y *Beauveria bassiana*. No menos importante se aprecia la investigación sobre infecciones y necrosis, hongos micorrícicos, biofertilizantes y nematodos, principalmente *Nacobbus aberrans;* también se destaca la investigación de enfermedades sobre chile *(Capsicum annuum)* como *Phytophthora capsici,* y en otros cultivos como aguacate, guayaba, papa, papaya, plátano y mango y sorgo *(Figura 8).*

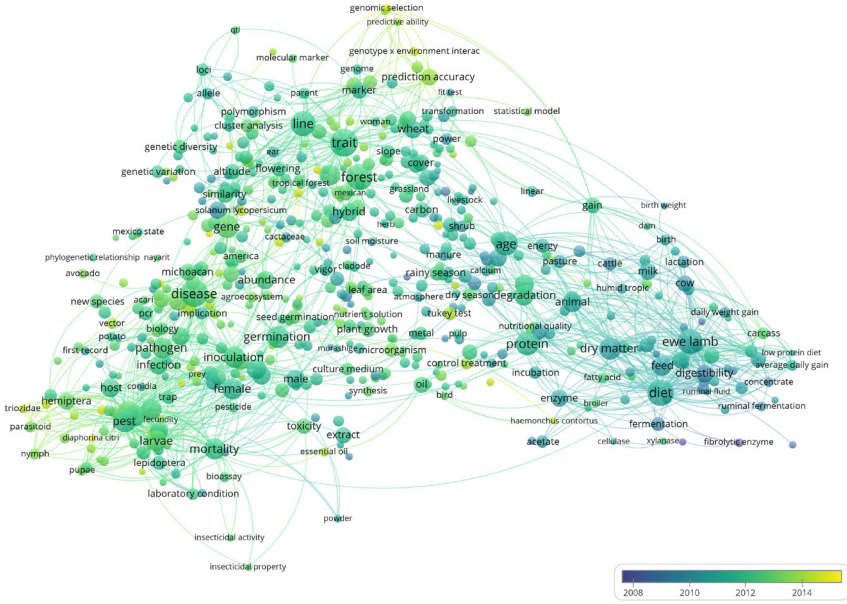
En los temas de investigación sobre genética vegetal (color morado), se destacan aquellos sobre características, genes y líneas de diferentes cultivos, híbridos (floración y llenado de grano) caracterización morfológica, mejoramiento genético, diversidad genética, variabilidad genética, selección de plantas, selección genómica, marcadores moleculares, polimorfismo, exactitud de la predicción, granos (principalmente trigo, maíz, haba y sorgo), y producción de semillas; en este *cluster* aparece también la investigación sobre cactáceas *(Figura 8).*

El *cluster* sobre cultivos agrícolas y bosques (color verde), muestra temas de investigación sobre bosques y su relación con el carbono en el suelo, producción maderera, especies y abundancia de árboles; se agrupan además en este *cluster* los temas sobre sistemas agrícolas, fertilidad de suelos, degradación de materia orgánica, pastizales, medio ambiente, y ecosistemas. Llama la atención que el tema sobre género se agrupa en este *cluster (Figura 8).*

El grupo de investigación en ciencia de las plantas (color azul) trata sobre diversos aspectos como el crecimiento de las plantas, la propagación, emergencia y la germinación de las semillas. Destacan, además, los temas sobre déficit de agua y salinidad; los minerales: potasio, fósforo, calcio, hierro y cobre; la actividad antioxidante, flavonoides, metabolitos, taninos y compuestos fenólicos; se agruparon aquí, además, temas como micorrizas, soluciones nutritivas, cultivo de tejidos, medios de cultivo, fitoremediación, extractos de plantas, aceites, actividad y propiedades insecticidas de plantas, conductividad eléctrica, prueba de Tukey, y aguas negras *(Figura 8).*

Los temas de investigación más actuales identificados (color amarillo en la *Figura 9* en los diferentes *clusters* mediante el análisis realizado son: agentes de control biológico, triózidos *(Triozidae),* psílido asiático de los cítricos *(Diaphorina citri),* psílido de la papa *(Bactericeria cockreli),* factores abióticos, chile, tomate, rendimiento y peso de frutos, alto rendimiento, híbridos comerciales, plantas medicinales, selección genómica, interacción genotipo ambiente, aceites esenciales, capacidad antioxidante, exactitud de la predicción, el parásito del estómago de rumiantes *Haemonchus contortus,* canal (de animales), prueba de Tukey y tratamiento de aguas residuales.

Figura 9. Temas de investigación más actuales (2014-2015) identificados mediante minería de textos de la investigación del Colegio de Postgraduados en revistas de corriente principal periodo 2005-2017.



CONCLUSIONES O DISCUSIÓN

La investigación en los primeros años del Colegio de postgraduados en revistas de corriente principal fue muy escasa, fue hasta 1972 cuando se publicaron los primeros documentos; hasta 1989, sólo se publicaron 157 documentos, 4.54% del total. Los principales temas sobre los que se publicó fueron: maíz y suelos; frijol, muy relacionado con la acumulación de la fitohormona ácido abscísico; hormigas, y el estudio de nuevas especies de interés, en ese tiempo para la investigación en ciencias agrícolas.

En el periodo de 1990 a 2004, la investigación empezó a tomar la forma que se muestra en los últimos años, se publicaron 580

documentos, que corresponden a 16.78% del total. Con tales registros, se formó un mapa donde se identificaron dos secciones con cuatro grupos o *clusters;* la primera sección está formada por el *cluster* de ganadería, y la restante sección está conformada por tres grupos o *clusters* con temas como producción de cosechas, suelos, y enfermedades y plagas de las plantas.

En el último periodo analizado, de 2005 a 2017, la investigación creció enormemente, y se llegaron a publicar en promedio casi 210 documentos por año, hecho que contrasta enormemente con el periodo 1972-1989, cuando sólo se publicaron ocho y medio documentos por año. Con estos registros, se formó un mapa con cinco *clusters* o grupos temáticos: ganadería, enfermedades y plagas de las plantas, genética vegetal, suelos y bosques, y ciencias de las plantas.

Entre los temas de investigación más actuales, se encuentran agentes de control biológico, triózidos y psílidos; algunos cultivos de gran consumo en México como, chile y tomate; híbridos comerciales, plantas medicinales, selección genómica, aceites esenciales, capacidad antioxidante, el parásito del estómago de rumiantes *Haemonchus contortus,* y el tratamiento de aguas residuales.

La minería de textos es una metodología que nos ayuda a encontrar información inmersa en los títulos y resúmenes de documentos como artículos científicos, que no están a simple vista. Esto abre una vía de investigación que ayuda a identificar las temáticas de investigación en documentos científicos; sin embargo, debe ser tratada con cuidado ya que no es una metodología sobre la que se tenga control del análisis de los documentos mediante los términos o palabras dentro de los registros, como sí sucede con las palabras clave o descriptores.

Se recomienda que, para tener una visión más completa de las temáticas de investigación del Colegio de Postgraduados, se realice un análisis de palabras conjuntas o co-palabras con las palabras clave de los registros bibliográficos proporcionadas por los autores de los documentos y por los indizadores de las bases de datos SCIE y SSCI.

BIBLIOGRAFÍA

Centre for Science and Technology Studies. «VOSviewer - Visualizing Scientific Landscapes». VOSviewer, 2018. http://www.vosviewer.com//.

Colegio de Postgraduados. «Colegio de Postgraduados». Conócenos, 2016. http://www.colpos.mx/wb/index.php/conocenos/.

———. «Línea de Tiempo». Conócenos, 2014. http://www.colpos.mx/wb/index.php/conocenos/linea-de-tiempo.

———. «Líneas de Generación y/o Aplicación del Conocimiento Institucionales». Investigación, 2018. http://www.colpos.mx/wb/index.php/investigacion/lineas-de-generacion-y-o-aplicacion-del-conocimiento-institucionales.

———. «Oferta Educativa». Educación, 2017. http://www.colpos.mx/wb/index.php/educacion/oferta-educativa.

Eck, Nees Jan Van, y Ludo Waltman. «Text mining and visualization using VOSviewer». *arXiv:1109.2058 [cs],* 9 de septiembre de 2011. http://arxiv.org/abs/1109.2058.

———. «VOS: A New Method for Visualizing Similarities Between Objects». En *Advances in Data Analysis,* editado por Reinhold Decker y Hans -J. Lenz, 299-306. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, 2007.

Feldman, Ronen, y James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press, 2006. https://doi.org/10.1017/CBO9780511546914.

Gobster, Paul H. «(Text) Mining the LANDscape: Themes and trends over 40 years of Landscape and Urban Planning». *Landscape and Urban Planning* 126 (1 de junio de 2014): 21-30. https://doi.org/10.1016/j.landurbplan.2014.02.025.

González Cossío, Félix. «Prólogo». En *Nuevas tendencias científicas y tecnológicas en el Colegio de Postgraduados,* 5-7. Montecillo, Texcoco, Estado de México: Colegio de Postgraduados, 2010.

Larqué-Saavedra, Alfonso. *Crónicas de la ciencia 2005-2014.* Mérida, Yucatán: CICY, Consejo Consultivo de Ciencias de la Presidencia de la Republica, 2014.

Salager-Meyer, Françoise. «Peripheral Scholarly Journals: From Locality to Globality». *Ibérica* 30 (1 de noviembre de 2015): 15-36.

Tan, Ah-Hwee. «Text Mining: The State of the Art and the Challenges». En *In Proceedings PAKDD'99 Workshopon Knowledge Discovery from Advanced Databases (KDAD'99),* 71-76, 1999. http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf.

Minería de texto aplicada a un diagnóstico de usuarios en Ciencia y Tecnología: aprendizajes para fortalecer la investigación bibliotecológica

Lourdes Feria Basurto
Consultora e investigadora independiente

Introducción

Las dos actividades de divulgación científica a la que menos asisten las familias en México son la *Semana Nacional de la Ciencia* y los *Talleres Itinerantes de Ciencia*, que ofrece el Consejo Nacional de Ciencia y Tecnología junto con sus contrapartes en los estados del país. La edición más reciente de la *Encuesta sobre la Percepción Pública de la Ciencia y la Tecnología* (ENPECYT) (INEGI 2015), preparada por el Instituto Nacional de Estadística y Geografía (INEGI) en conjunto con el Consejo Nacional de Ciencia y Tecnología (Conacyt), identifica como uno de los indicadores de involucramiento en esos temas, por parte de la ciudadanía, el porcentaje de visitas a recintos y actividades vinculadas con la apropiación del conocimiento, e informa que a nivel nacional la *Semana Nacional de Ciencia y Tecnología,* es la que se reporta como la opción menos favorecida en cuanto a asistencia, lo que deja en antepenúltimo y penúltimo sitios las exposiciones tecnológicas

e industriales y los planetarios, que se ven rebasados ampliamente por la asistencia a los cines, parques de diversiones y zoológicos/acuarios.

Ilustración 1: Población que visitó diferentes tipos de recintos

Indicador	2013	2015
Zoológico o acuario	42.2	31.0
Biblioteca pública	24.1	23.0
Museo de ciencia y tecnología	16.3	17.8
Planetario	12.9	12.3
Exposiciones tecnológicas o industriales	18.5	12.6
Semana nacional de ciencia y tecnología	8.2	7.8
Museo (de arte, de cera, natural)	26.4	26.4
Parque de diversión	49.6	38.9
Teatro	22.9	19.1
Cine	NA	55.7

Notas y Llamadas:
/a Población de 18 años y más.
NA No aplica
Los valores no son sumables, dado que se trata de una pregunta de opción múltiple.
Los valores pueden variar debido al redondeo.
Fuente:
INEGI. CONACYT Instituto Nacional de Estadística y Geografía. Consejo Nacional de Ciencia y Tecnología. Encuesta sobre la Percepción Pública de la Ciencia y la Tecnología (ENPECYT) 2013, 2015

Ante esa realidad, en el estado de Colima, el Consejo Estatal de Ciencia y Tecnología (Cecytcol) instrumentó en 2017 un estudio cuyo objetivo era conocer el impacto de la apropiación social de la ciencia en todos los municipios del estado, en las comunidades y en las escuelas de los niveles primaria, secundaria y bachillerato. Para desarrollarlo se partió de una revisión documental y archivística de los últimos tres años fiscales (2014, 2015 y 2016), así como de un levantamiento de datos *in situ* durante los meses de octubre 2017 a abril 2018, a fin de reconocer las áreas de oportunidad

que tienen las actividades de divulgación en la entidad para, en lo futuro, mejorarlas buscando la congruencia con su *Plan estatal de desarrollo* (Colima 2016), que propone la construcción de una economía del conocimiento con mayores oportunidades para los jóvenes, partiendo de una realidad que muestra la persistencia del rezago educativo, una cobertura insuficiente y una baja calidad en la educación, pero con la mirada puesta en impulsar una política a favor de la innovación, el fortalecimiento del vínculo sector productivo - generación de conocimiento, la mejora de la conectividad del estado, así como la reducción de las brechas educacionales y la armonización de la educación con las necesidades del mercado laboral; haciendo énfasis en uno de sus objetivos (II.3.4.1.2) y "ampliar la divulgación de la ciencia y la tecnología en los niveles medio superior y superior" (Colima 2016, 115).

Discursos y narrativas como fuentes de datos

Los insumos informacionales que permitieron obtener testimonios orales en la forma de discursos, historias de vida y narrativas partieron del planteamiento de la pregunta clave que guió el estudio: ¿cómo atraer a más personas a actividades de información y conocimiento?, esto se resolvió estructurando una metodología mixta para el levantamiento de datos, que comprendió seis etapas:

> Etapa 1: Investigación documental y archivística.
> Etapa 2: Observación participante e involucramiento con las comunidades.
> Etapa 3: Etnofotografía y diarios de campo de investigación acción.
> Etapa 4: Encuestas a estudiantes asistentes a los talleres (aplicación de 381 cuestionarios a estudiantes de nivel básico, medio y medio superior).
> Etapa 5: Grupos focales con profesores y con divulgadores de la ciencia.

Etapa 6: Entrevistas a profundidad con profesores y divulgadores de la ciencia.

Para los fines de la presente revisión, se hará énfasis en las etapas 2, 3, 5 y 6 y se describirán a continuación las técnicas aplicadas en cada una de ellas.

Observación participante e involucramiento con las comunidades (Etapa 2)

El trabajo etnográfico comenzó con la observación sistemática y el levantamiento de notas de campo durante catorce semanas en las que se registraron los eventos significativos de cada día junto con las interpretaciones de los informantes. Las observaciones iniciales se centraron en la recopilación de datos generales y abiertos. Este proceso fue importante para recabar antecedentes para una investigación más centrada y también para establecer una buena relación con los informantes, evitar interpretaciones parciales y probar si las preguntas de investigación originales resultaban significativas y pertinentes.

Por otra parte, se realizó una intervención dentro de las actividades de divulgación como asistentes/oyentes entre las personas estudiadas durante un periodo de seis meses, se recopilaron datos mediante la participación continua en los talleres, charlas, etcétera.

Etnofotografía y diarios de campo de investigación-acción (Etapa 3)

Además de las observaciones escritas, los registros y las bitácoras, la investigación cualitativa se apoyó en levantamientos etnofotográficos en imagen fija y video, lo que integró una galería de más de novecientas fotografías y dieciséis videos y audiograbaciones. Como parte de las actividades de investigación-acción dos de los integrantes del grupo de investigación formaron parte activa al integrarse como conferenciantes en la modalidad de "Charla con un

Científico" e impartir en tres diferentes locaciones rurales la conferencia denominada "Los drones y tú", evento que generó el valor agregado de observar una atmósfera de valoración favorable hacia la ciencia y el interés de los asistentes, en su mayoría niños entre los siete y doce años de edad.

Grupos focales y entrevistas a profundidad con profesores y divulgadores de la ciencia (Etapas 5 y 6)

Después de la orientación inicial, se siguió un programa sistemático de entrevistas formales con base en una batería de cuestionamientos relacionados con las preguntas de investigación. Para ello se seleccionaron veintiún informantes clave para investigar los patrones de percepciones. A partir de ese universo, se hicieron catorce entrevistas a profundidad y dos sesiones de grupos focales. La selección de informantes clave se realizó mediante la variante de *muestreo de juicio* cuidando elegir sujetos bien informados, confiables y que pudiesen informar de los datos contextuales y reconocer los elementos significativos así como las interconexiones a medida que se desarrollaban las secuencias de entrevistas. Desde la perspectiva del análisis del impacto se consideraron, en primer lugar, los elementos simbólicos y se registraron observaciones con la debida atención tanto al contexto cultural como a los significados asignados por los involucrados.

Asimismo, con el fin de dar mayor sustento a esta vertiente de la investigación cualitativa, se hizo previamente una revisión cuantitativa de los informes 2014-2016, así como un reporte de talleristas a partir de lo cual se pudieron extraer inferencias validadas estadísticamente.

Metodología para el manejo de datos: experimentación con minería de texto

Las cuatro etapas descritas permitieron recabar fuentes de datos primarias sobre las percepciones ciudadanas, así como los comportamientos y expresiones individuales hacia la divulgación científica. Con ello se produjo una base de conocimiento integrada por documentos fotográficos, informes, entrevistas, conversaciones y las correspondientes notas de trabajo de campo basadas en la observación sistemática realizada durante catorce semanas, cuya evidencia quedó registrada en dieciséis expedientes de transcripciones basadas en audio y videograbaciones a partir de dos grupos focales y catorce entrevistas; un catálogo/bitácora de cerca de mil piezas de fotografía etnográfica catalogadas y analizadas, un diario de campo incluyendo notas de campo semanales y reportes de observación participante, y un archivo digital de cuatrocientos párrafos testimoniales todo lo cual hizo posible identificar unidades de valor para su filtrado y análisis.

Tras el levantamiento de datos cualitativos se trabajó la información mediante minería de texto, técnica que ha sido descrita como

> […] un campo interdisciplinario que combina técnicas de lingüística, computación y estadística para recuperar y extraer información a partir de texto digital (Bergman, Hunter y Rzhetsky 2013, 210); y también como un proceso automatizado para grandes cantidades de datos textuales, no estructurados, para la recuperación, extracción, interpretación y análisis de información (Reilly 2012).

Otros términos con los que se conoce a la minería de texto son: minería de información, arqueología de información, gestión de conocimiento, data mining, etc., dependiendo del autor pero a lo que nos lleva es a que surja "la necesidad de darle un valor adicional a la información documental" (Justicia de la Torre 2017, 2).

Minería de texto tiene que ver con datos textuales no estructurados y el objetivo es que mediante la aplicación de algoritmos de minería informática se transforme la información textual en

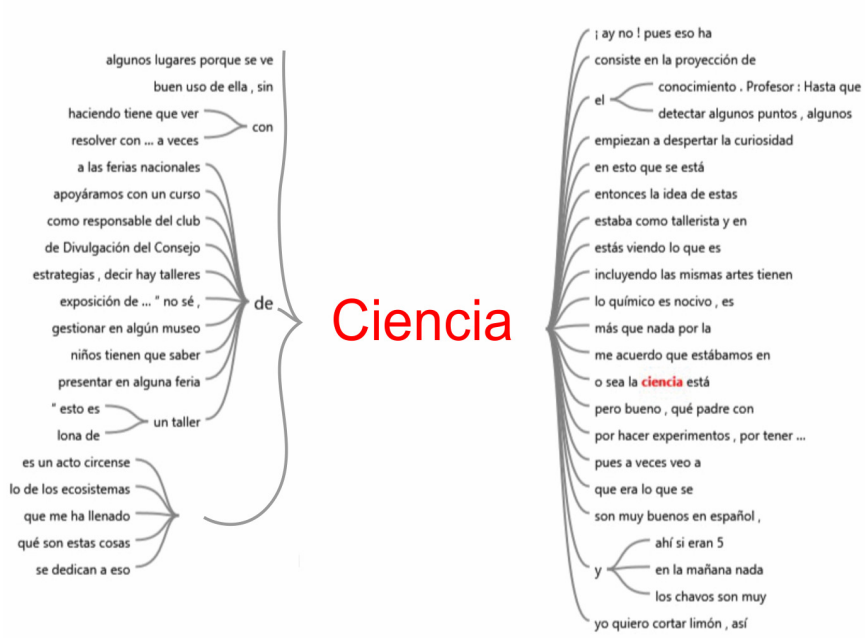
números y pueda identificarse nuevo conocimiento. En síntesis se trata de la aplicación de algoritmos informáticos al texto de las entrevistas; de tal suerte que a partir del lenguaje coloquial no estructurado se generen datos numéricos, vectores e indicadores; lo cual, expresado en términos matemáticos sería: la cuádrupla [DQFR(qi,dj)] donde "D" es un conjunto de vistas lógicas de documentos; "Q" es un conjunto de consultas de usuario; "F"es el marco de trabajo que vamos a usar para modelar y "R(qi,dj)" correspondería a la función *ranking* (Justicia de la Torre 2017).

En el caso particular de los datos cualitativos arrojados, éstos se procesaron mediante un *software* especializado que, en una primera etapa, permitió identificar patrones semánticos a través de los datos para hacer contrastaciones de las distintas historias, para posteriormente estructurar un mapeo semántico que permitió identificar dentro del *corpus* de textos la coocurrencia de las distintas palabras. Para ambos fines se digitalizaron las catorce entrevistas y los dos reportes de grupos focales para integrarlos en la aplicación digital *NVivo 11;* herramienta que permitió, por una parte, entender la frecuencia y asociaciones terminológicas para armar mapas y redes terminológicas; y por la otra, representar cada término en un espacio vectorial donde aquellas palabras con significado similar lograban estar más cerca en el trazado para hacer cálculos de frecuencias y porcentajes. Algunos de ellos se ilustran a continuación.

I*lustración 2:* Porcentaje ponderado de palabras en narrativas

Palabra	Longitud	Conteo	Porcentaje ponderado %
ciencia	5	80	1.16
niños	7	71	1.03
talleres	8	70	1.01
actividades	11	49	0.71
profesor	8	33	0.48
escuelas	8	30	0.43
trabajar	8	28	0.40
taller	6	27	0.39

Ilustración 3: Árboles semánticos



Los resultados de la investigación reflejan que, en general, lo que piensa la comunidad en Colima es que la ciencia es agradable y que cuando se les da a conocer a los niños y a los jóvenes se logra entusiasmarlos sinceramente. Por otra parte, se observa que la difusión sólo permea en las escuelas; que existe articulación entre talleres y programas de estudio; que es necesaria mayor actividad en zonas rurales, y que es necesario sensibilizar a las autoridades y motivar a los padres de familia.

Tabla 1. Tabla de resultados

Porcentaje	Concepto
75%	La Ciencia agradable es posible en talleres.
75%	La difusión sólo permea en escuelas.
75%	Necesaria mayor actividad en zona rural.
69%	Se logra motivar la vocación científica.
44%	Necesario sensibilizar autoridades.
38%	Existe articulación entre talleres y escuela.
31%	Necesario motivar padres de familia.

Resultados

Cabe señalar que previamente a la aplicación del *software* se llevó a cabo un análisis general y una limpieza de datos para proceder al ingreso de la información a fin de producir las tablas porcentuales de palabras más frecuentes, incluyendo la longitud de cada término, el dato sobre la cantidad de ocurrencias en el texto y el porcentaje ponderado de aparición. También se elaboraron tanto conglomerados a partir de palabras, como árboles semánticos que permitieron ver toda la pre y post-coordinación a partir del metadato clave que se eligió como elemento base; con todo ello se pasó a la fase de discernimiento a partir de la observación de vectores de coincidencias.

Conclusiones

Al final del estudio, las conclusiones emanadas se integraron en tres enunciados:

De textos a números y porcentajes. Fue posible a partir de textos no estructurados obtener formas intermedias numéricas que permitieron rescatar y medir aspectos relevantes y de ahí generar

nuevo conocimiento. De haber trabajado en forma manual, no necesariamente se hubieran podido identificar, o el tiempo para lograrlo habría sido mucho mayor.

No todo es inteligencia artificial. Se requiere de la Intervención humana para la limpieza de datos, la integración y la selección de los mismos. Todas las aplicaciones de minería tienen que ver con la participación del investigador y sus colaboradores, quienes hacen posible que el *software* ejecute de manera precisa las funciones necesarias.

Bibliotecas y manejo de datos. Las herramientas, las técnicas, el almacenamiento de datos, la recuperación y los métodos analíticos aún están en proceso de evolución, pero cada vez más las bibliotecas tendrán que fortalecerse en el uso de estos métodos y técnicas para orientar a los investigadores en sus proyectos. Entonces, ¿por qué no se va convirtiendo la biblioteca en el laboratorio natural para la gestión y organización de datos, así como en el área que se haga cargo de la capacitación permanente sobre la alfabetización en datos?

Bibliografía

"Research on tacit knowledge mining of university libraries based on data mining." 13Th International Conference On Service Systems And Service Management (ICSSSM), Service Systems And Service Management (ICSSSM), 2016 13Th International Conference On 1. *IEEE Xplore Digital Library,* 2016.

Botta Ferret E, Cabrera Gato JE. "Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital". *Acimed* 16, no. 4 (2007).

Bergman, Hunter y Rzhetsky (2013) citado por Dyas-Correia, Sharon, and Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Bernard Reilly (2012) citado por Dyas-Correia, Sharon y Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Cleary P, Garlock K, Novak D, Pullman E, Mann S. "Text Mining 101: What You Should Know. *Serials Librarian.* January 2017;72(1-4):156-159.

Connaway, Lynn y Marie L. Radford. *Research methods in Library and Information Science.* 6a. ed. Santa Barbara, CA, Libraries Unlimited, 2017.

Connaway, Lynn, S., William Harvey, Vanessa Kitzie, y Stephanie Mikitish. *Academic Library Impact: Improving Practice and Essential Areas to Research.* Chicago: Association of College and Research Libraries, OCLC Research, 2017.

Consejo Estatal de Ciencia y Tecnología del Estado de Colima, Consejo Nacional de Ciencia y Tecnología y Gobierno del Estado de Colima. *Estrategia nacional para fomentar y fortalecer la difusión y divulgación de la ciencia, la tecnología y la innovación en las entidades federativas: Colima.* Colima: CECYTCOL, 2014. Trabajo presentado en 21ª Semana Nacional de Ciencia y Tecnología. (Recuperado de: 21SNCT-COLIMA.docxs)

Contreras Barrera, Marcial. Minería de texto: una vision actual. *Bibl. Univ.,* 17, no. 2 (2014), 129-138.

Dyas-Correia, Sharon, and Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Faniel, Ixchel y Lynn S. "Librarians' Perspectives on the Factors Influencing Research Data Management Programs". *College & Research Libraries Journal: 79,* num. 1, (2018).

Instituto Nacional de Estadística y Geografía. Encuesta sobre la percepción pública de la Ciencia y la Tecnología (ENPECYT). México, INEGI, CONACYT, 2015

Justicia de la Torre, María Consuelo. *Nuevas técnicas de minería de textos: aplicaciones.* Granada: Universidad de Granada, 2017.

Mariñelarena-Dondena, Luciana, Marcelo Luis Errecalde y Alejandro Castro Solano. "Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología" *Revista Argentina de Ciencias del Comportamiento,* 9, no. 2 (2017), 65- 76.

Morris, Walter. "Text Mining for the Social Sciences" *Cornerstone 3 Reports: Interdisciplinary Informatics.* Paper 53 (2011) Santana Mansilla, Pablo; Costaguta, Rossana y Daniela Missio. "Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos". *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial.* 17, no. 53, (2014), 57-67.

Yu, C. H., Jannasch-Pennell, A., y DiGangi, S. "Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability". *The Qualitative Report,* 16, no. 3, (2011), 730-744. http://nsuworks.nova.edu/tqr/vol16/iss3/6

Minería de Datos, el caso de estudio de la Biblioteca Dr. Valentín Gómez Farías de la Facultad de Medicina de la UNAM.

David Flores Macías
Guadalupe Vanesa Carolina Gutiérrez Hernández
Universidad Nacional Autónoma de México

Introducción

La Minería de Datos es el proceso automatizado para la extracción de patrones de un cierto conjunto de datos. Aunque es éste un paso del Proceso de Descubrimiento de Conocimiento, normalmente se le conoce como Minería de Datos. También se puede definir como el hecho de descubrir información implícita pero útil de datos almacenados.

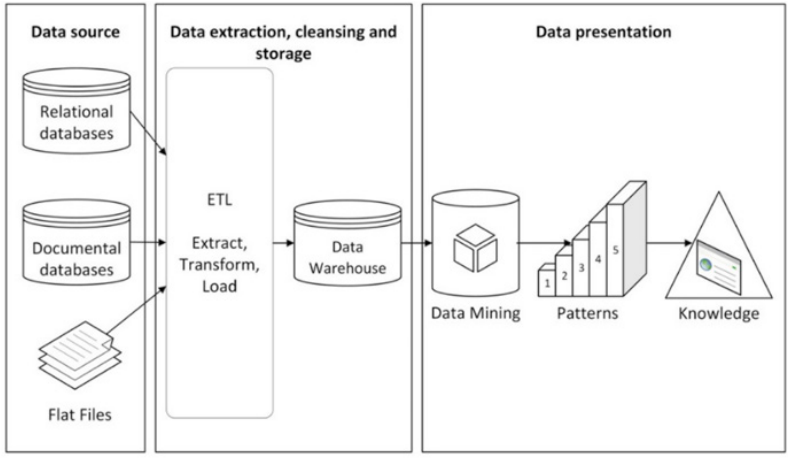
Las técnicas de minado son aplicadas en un amplio rango de dominios; por mencionar algunos ejemplos, si se genera minería de datos con datos obtenidos de la web, se conoce como *web mining,* si es usado en texto es minería de textos y si se aplica a Bibliotecas es llamado *Bibliomining* o Bibliominería. Este último término es muy interesante debido a que si uno realiza la búsqueda en inglés en la web de bibliotecas y minería de datos, normalmente los resultados proporcionan información técnica sobre las librerías utilizadas por los algoritmos de máquina. Por ello

(Nicholson 2006), se introdujo el término de Bibliomining, justamente para hacer referencia a la aplicación de la minería de datos en Bibliotecas. Siendo más específicos, en el presente trabajo la Bibliominería es usada para encontrar patrones y tendencias de los sistemas transaccionales en bibliotecas, entendiéndose como transaccionales todas aquellas operaciones que se realizan en una base de datos al realizar movimientos de circulación tales como préstamos, devoluciones y resellos (Prakash *et al.* 2004).

Desarrollo

El proceso de Minería de Datos utilizado en este estudio se presenta a continuación (Sigüenza-Guzmán 2015):

Diapositiva 1



1. Origen de los datos. Tomando en cuenta la estructura de la base de datos de circulación bibliográfica, se identificaron aquellos campos de la misma que podrían ser útiles para el estudio, y que

también fueran candidatos para poderse categorizar y construir la vista minable. Se determinó que éstos fueran la carrera del alumno, el material bibliográfico la clasificación, la fecha de préstamo, la fecha de devolución (indicada por sistema), la fecha de retorno (fecha real en la que se realizó la devolución) y la hora del préstamo.

2. Extracción de los datos, limpieza y almacenamiento.
Creación de una vista minable (Gutiérrez, Barranco y Méndez 2008). Para obtener dichos datos, se ejecutó una consulta SQL en el Sistema Manejador de Bases de Datos Oracle. El periodo fue del 1-08-2015 al 31-10-2018, dicha consulta proporcionó un total de 133 776 registros.

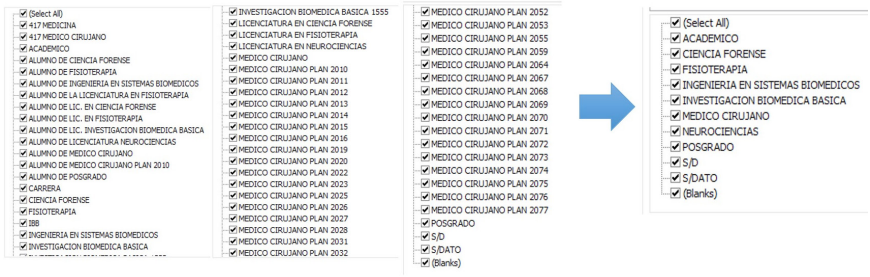
Diapositiva 2

CLASIFICACION	FECHA PR	FECHA DE	FECHA RE	HORA PRESTAMO	CARRERA
RE461 K3518 2016	20180611	20181113	0	1827	MEDICO CIRUJANO PLAN 2010
RE461 K3518 2016	20180611	20181113	0	1826	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181107	0	1539	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181106	0	1516	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180724	20181115	0	1725	MEDICO CIRUJANO
RM300 B3618 2016	20180813	20181107	0	906	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181106	0	854	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20180827	0	851	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180522	20180605	0	1507	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180821	20181115	0	859	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180824	20181017	0	1319	MEDICO CIRUJANO PLAN 2010
RC731 C355 2017	20180724	20181115	0	1725	MEDICO CIRUJANO
RC111 M35 2016	20180829	20181108	0	1411	MEDICO CIRUJANO
RC76 M6818 2015	20180828	20181105	0	1544	MEDICO CIRUJANO
RC76 M6818 2015	20180820	20181029	0	1415	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180629	20180803	0	1111	MEDICO CIRUJANO
RC76 M6818 2015	20180821	20181015	0	1446	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180730	20181106	0	1433	ALUMNO DE MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180825	20181112	0	935	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180823	20181026	0	1705	MEDICO CIRUJANO PLAN 2010
RA972 G65 2016	20180615	20181113	0	1522	MEDICO CIRUJANO PLAN 2010
R899 H4718 2016	20180628	20181110	0	1722	MEDICO CIRUJANO PLAN 2010
R899 H4718 2016	20180822	20181113	0	1355	MEDICO CIRUJANO
R899 H4718 2016	20180725	20181114	0	1508	ALUMNO DE MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181107	0	1258	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	851	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	1201	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	1140	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20180827	0	1020	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	943	MEDICO CIRUJANO PLAN 2010

A continuación se muestra cómo se limpiaron o categorizaron cada uno de los campos seleccionados.
Campo: Carrera del alumno. Limpieza de los datos.

Como se puede ver en la figura, el campo de carrera no se encontraba normalizado, y existían diversas formas de nombrar una misma carrera. En este caso, utilizando el *software* de aplicación Excel, a través de fórmulas, se realizó la categorización, la cual quedó de la siguiente forma.

Diapositiva 3



Campo: Clasificación. Categorización de los datos.
A través de funciones de Excel, y de acuerdo con cada una de las clasificaciones de los registros, se recuperó el nombre de la clase o materia, de acuerdo con la clasificación LC (Library of Congress).

Diapositiva 4

CLASIFICACION	Clasificacion LC	Nombre
RE461 K3518 2016	RE	Ophthalmology
RE461 K3518 2016	RE	Ophthalmology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RC731 C355 2017	RC	Internal Medicine
RC111 M35 2016	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RA972 G65 2016	RA	Public Aspects of Medicine
R899 H4718 2016	R	Medicine (General)
R899 H4718 2016	R	Medicine (General)
R899 H4718 2016	R	Medicine (General)
QM601 M65418 2016	QM	Human Anatomy

Campo: Estatus de préstamo. Categorización de los datos.
Se categorizó de la siguiente forma:

PT. Libros que se encuentran prestados en tiempo
PV. Libros prestados que no han sido devueltos.
DT. Devoluciones realizadas en tiempo.
DV. Devoluciones realizadas después de la fecha de devolución indicada en el sistema.

Campo: Hora de préstamo. Categorización de los datos.
Si la hora se encuentra dentro del rango de 8:00 a 15:00, se estableció como TM (Turno matutino).
Si la hora se encuentra dentro del rango de 15:01 a 20:00, se estableció como TV (Turno vespertino).
Con todos estos campos limpios y categorizados, fue posible obtener la vista minable, de la cual se muestra a continuación un extracto.

Diapositiva 5

Clasificacion	Nombre	Estatus	Carrera	HORA
RE	Ophthalmology	PT	MEDICO CIRUJANO	TV
RE	Ophthalmology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TV
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TV
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TV

Esta vista minable se exportó de Excel a un archivo delimitado por comas, el cual fue el archivo de entrada para el *software* de aplicación que se encargó de realizar el minado de datos.

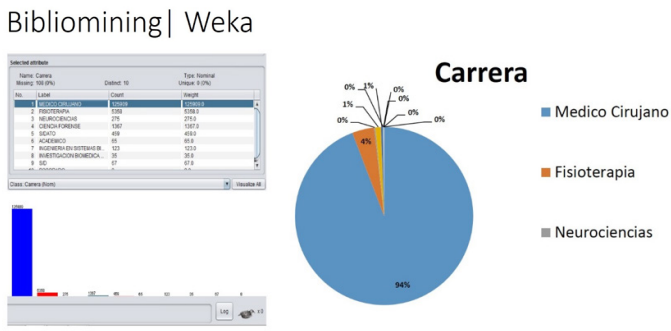
3. Minería de datos para generar conocimiento y presentación de los datos.
Con la vista minable ya generada, se decidió realizar las tareas de *Clustering* (Agrupación) y Clasificación, con el fin de encontrar patrones no triviales.

3.A *Clustering*
Objetivo: se identificaron grupos de registros que son similares entre ellos, pero diferentes del resto de los datos.

Software utilizado: Weka (Weka 3) es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, su clasificación, regresión, agrupación, extracción de reglas de asociación y visualización. Es un *software* de código abierto emitido bajo la Licencia Pública General de GNU.

Weka proporciona un primer vistazo estadístico de los datos contenidos en la vista minable.

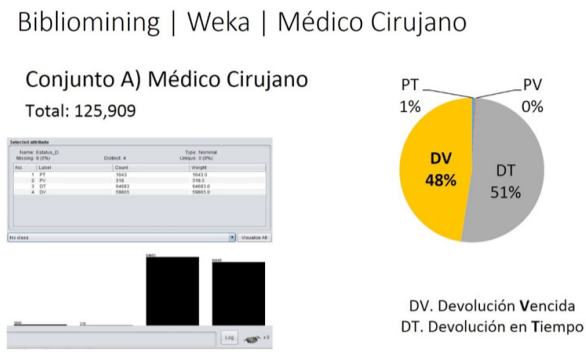
Diapositiva 6



Esta imagen nos indica que el 94% de los datos pertenecen a la carrera de Médico Cirujano, por lo que se decidió dividir el conjunto en dos apartados, lo que quedó de la siguiente forma:

Conjunto A.1) Médico cirujano (125 909 registros)
Conjunto B.1) Otras carreras y clasificaciones.
Comenzado con el Conjunto A.1) Médico cirujano.

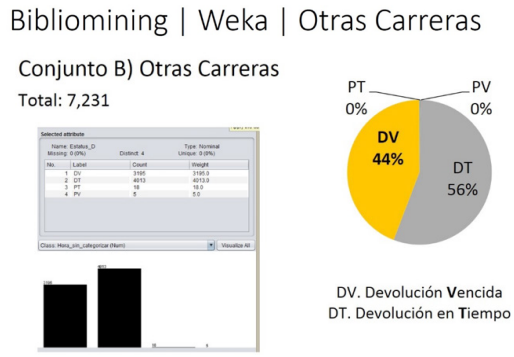
Diapositiva 7



Esta gráfica nos indica que el 99% de los libros que se prestan, son devueltos a la biblioteca, pero de ellos, el 51% se regresa de manera tardía; es decir, después de la fecha indicada en el sistema.

Un fenómeno muy parecido sucede con el conjunto B.1, de las otras carreras.

Diapositiva 8



Tomando las devoluciones vencidas, se decidió utilizar dicho campo como base para implementar la tarea de *Clustering* con el fin de identificar grupos de registros que son similares entre ellos, pero diferentes del resto de los datos.

Los resultados para el grupo A.1 (Médico Cirujano), fueron los siguientes:

Diapositiva 9

```
Final cluster centroids:
                                                Cluster#
Attribute                                Full Data          0          1          2          3
                                         (125909.0) (43046.0) (35589.0) (40945.0) (6329.0)
=============================================================================================
Materia                                  Human Anatomy Internal Medicine     Human Anatomy     Human Anatomy Internal Medicine
Estatus_D                                           DT                DT                DT                DV                DT
Carrera                                  MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO
Hora_Prestamo                                       TM                TV                TM                TM                TM



Time taken to build model (full training data) : 0.48 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       43046 ( 34%)
1       35589 ( 28%)
2       40945 ( 33%)
3        6329 (  5%)
```

El libro que corresponde a la clasificación de Anatomía Humana, que normalmente se presta en el turno matutino, tiende a devolverse de manera tardía.

Con relación al grupo de otras carreras, los resultados proporcionados por la herramienta fueron:

Diapositiva 10

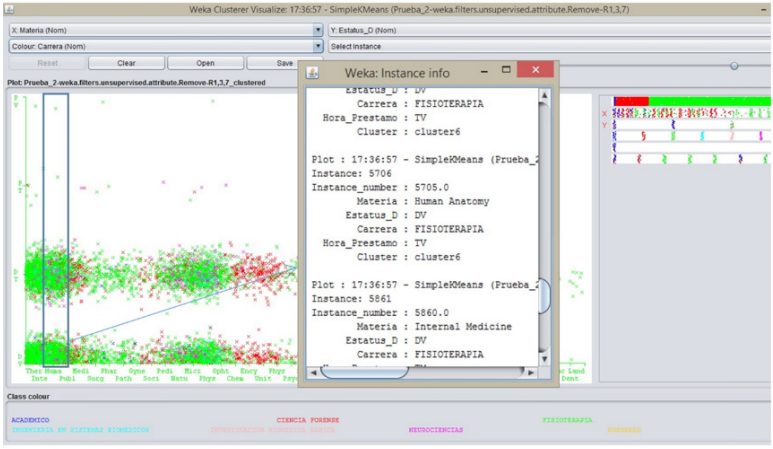
Bibliomining | Weka | Otras Carreras

```
                                5                   6                   7
                          (611.0)             (1958.0)             (137.0)
=============================================================================
Public Aspects of Medicine              Human Anatomy   Natural History, Biology
                       DV                          DV                        DV
                CIENCIA FORENSE         FISIOTERAPIA        FISIOTERAPIA
```

Lo que esto nos indica es que los alumnos de las carreras de Ciencia Forense que obtienen los libros de aspectos públicos de la medicina y los alumnos de Fisioterapia que se llevan en préstamo los libros con clasificación de Anatomía Humana, Historia Humana y Biología, representan a aquellos que devuelven los libros de manera tardía.

Adicionalmente WEKA nos muestra de manera gráfica, cómo es que se visualizan los datos; aquí el ejemplo para el conjunto B, de otras carreras.

Diapositiva 11



3.B Clasificación

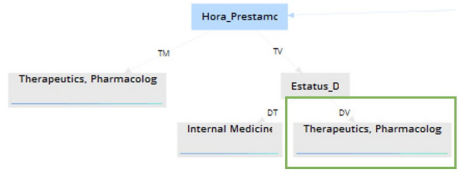
Al ser éste un aprendizaje supervisado, asigna elementos de una colección a categorías o clases de destino.

Software utilizado: RapidMiner es un programa para realizar minería de datos. No es *software* libre, cuenta con una versión educativa.

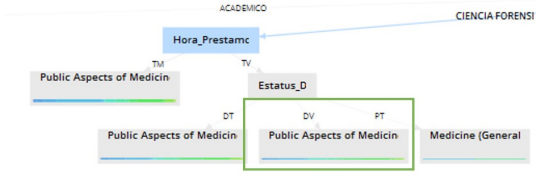
RapidMiner, con la ayuda del asistente, de manera muy rápida, permite establecer la tarea de minería de datos. Siguiendo los pasos del asistente y seleccionando la tarea de clasificación, es posible obtener arboles de decisión, que presentan información de cada una de las carreras.

Diapositiva 12

Bibliomining | rapid miner | Otras Carreras



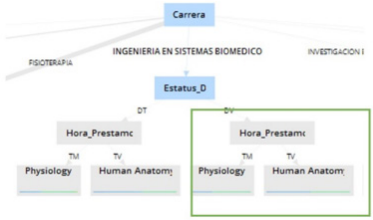
Académico



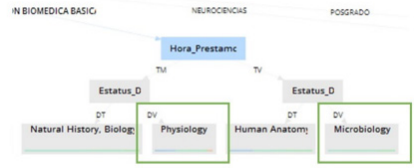
Ciencia Forense

Diapositiva 13

Bibliomining | rapid miner | Otras Carreras

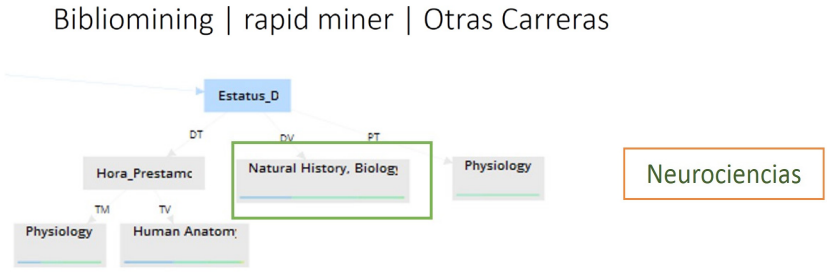


Ingeniería en Sistemas Biomédicos



Investigación Biomédica Básica

Diapositiva 14



Recopilando toda esta información, se puede resumir el conocimiento generado relacionado con el comportamiento de devoluciones tardías.

Conclusiones

Con el apoyo de la estadística, se detectó que un gran porcentaje de los libros que se prestan y que son devueltos, lo son de manera tardía (DV) (44%-46%).

Aplicando tareas de minería de datos, es posible conocer de dichas devoluciones vencidas, a qué clasificación pertenecen y en qué horario fueron prestadas.

Con dicha información se podría establecer que la multa no es factor importante para la devolución del material bibliográfico; se tendría que revisar la política para mejorar el regreso de libros en tiempo.

El hecho de conocer la clasificación de los libros que se devuelven de manera tardía, motiva a realizar nuevos análisis de estudio de la colección, poniendo atención en dichas clasificaciones.

Fuentes consultadas

Bin, Chen. 2013. "Study on Data Mining in Digital Libraries." In , 282–91. *Springer, Berlin, Heidelberg.* https://doi.org/10.1007/978-3-642-53703-5_30.

Candás Romero, Jorge. 2006. "Minería de datos en bibliotecas: bibliominería." 2006. http://bid.ub.edu/17canda2.htm.

Nicholson, Scott. 2006. "The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services." *Information Processing & Management* 42 (3): 785–804. https://doi.org/10.1016/j.ipm.2005.05.008.

Sarma, Pankaj Kumar Deva, and Rahul Roy. 2010. "A Data Warehouse for Mining Usage Pattern in Library Transaction Data." Assam University Journal of Science and Technology. http://www.inflibnet.ac.in/ojs/index.php/AUJSAT/article/view/194.

Zhu, Tingting, and Lili Zhang. 2011. "Application of Data Mining in the Analysis of Needs of University Library Users." 2011 6th International Conference on Computer Science & Education (ICCSE), Computer Science & Education (ICCSE), 2011 6th International Conference On. https://doi.org/10.1109/ICCSE.2011.6028662.

Juan Camilo Giraldo, Mejía, and Builes Jovani Alberto Jiménez. "Caracterización del Proceso de Obtención de Conocimiento y Algunas Metodologías para Crear Proyectos de Minería de Datos." *Revista Latinoamericana de Ingeniería de Software,* Vol 1, Iss 2, Pp 42-44 (2013) no. 2 (2013): 42. Directory of Open Access Journals, EBSCOhost (accessed September 7, 2018).

Gutiérrez Hernández, Guadalupe Vanessa Carolina, Verónica Barranco Serrano, and Carlos Francisco Méndez Cruz. Minería de datos dentro del proceso de KDD aplicado a la base de datos de circulación bibliográfica de la Biblioteca Central. n.p.: 2008. TESIUNAM, EBSCOhost (accessed September 7, 2018).

Prakash, K & Chand, Prem & Gohel, Umesh. (2004). Application of Data Mining in Library and Information Services. Presented at the 2nd Convention PLANNER, Manipur Uni., Imphal.

Weka 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/

RapidMiner. Lightning Fast Data Science for Teams. https://rapidminer.com/

SISTEMATIZACIÓN DE DATOS Y SERVICIOS DE INFORMACIÓN

Research Data Management and Libraries: Opportunities and Challenges

KRYSTYNA K. MATUSIAK
University of Denver

INTRODUCTION

Research Data Management (RDM) is a new area of service and infrastructure development at universities and research centers worldwide. The increasing volume and complexity of digital data, as well as the challenges associated with organization, preservation, and reuse of data, have contributed to the emergence of RDM as a priority in recent years. Modern science has increasingly become data-intensive with researchers using new methodology and instruments and producing an unprecedented amount of data (Borgman 2012). Digital technology has accelerated this process by providing new tools for collecting scientific evidence but also enabled building technical infrastructure for storing and sharing data. The researchers studying the growth of science found that global scientific output doubles every 9 years. Most of the scientific expansion has taken place in the modern era with the growth rate of 8 to 9% (Bornmann & Mutz 2015).

The motivations for deployment of RDM services are diverse, often emerging from a pragmatic need to comply with requests from funding agencies for data management planning, but also responding to the policy environment and calls for openness in science (Ayris *et al.* 2016; Fearon et al 2013; Pryor *et al.* 2013). National funding agencies in several countries now require researchers to prepare data management plans and to provide open access to data (NSF; UK Research and Innovation). The European Research Council (ERC) supports the principle of open access to research data and scholarly publications. It conducted a Pilot on Open Research Data for research projects funded through the Horizon 2020 program. As of 2017, the Pilot on Open Research Data has been extended and open access became the default for the research data generated as a result of the Horizon 2020 funding, although researchers can still opt out in some circumstances (ERC 2018). In addition to funder requirements, journal editors and publishers are increasingly requesting authors to provide open access to source data underpinning publications.

This paper provides an overview of RDM services and their importance in the context of Open Science. It summarizes the findings from the Data Curation project sponsored by the International Federation of Library Associations (IFLA) Library Theory and Research (LTR) Section. The IFLA study focused on the roles and responsibilities of RDM professionals in international and interdisciplinary contexts. This paper discusses the opportunities and challenges in providing RDM services in light of the findings from the IFLA Data Curation project.

Open data and the open science movement

In the traditional scholarly communication model, scholars disseminated the results of their research through conference presentations, books, and articles published in peer-review, subscription-based journals. The Open Access (OA) movement has changed the model of scholarly publishing encouraging scholars

to share their papers through open access publishing or depositing published articles in institutional or disciplinary repositories (Swan 2012). The emphasis of OA, however, has been almost exclusively on opening access to journal articles, not so much on research data. As Borgman (2015) notes open data is "substantially distinct from open access to scholarly literature" (p. 44). Researchers would sometimes share data sets with colleagues in the scholarly community but rarely provide open access as part of the traditional scholarly communication practice.

Data is a valuable output of scholarly work and the calls for providing open access to research data come not only from the funding agencies but also from the members of the scholarly community. Opening access to data is believed to contribute to transparency and reproducibility of research and to the more efficient scientific process (Kraker *et al.* 2011; Molloy 2011; Nosek *et al.* 2015). Open research data can be freely accessed, reused, and redistributed for scholarly purposes. The principles of FAIR data (findable, accessible, interoperable and reusable) provide a foundation for access and reuse of research data across disciplines and borders (Wilkinson *et al.* 2016). Open Data is a key component of the Open Science movement.

The Open Science movement advocates for opening all phases of the research cycle and sharing all outcomes of the scientific work (Foster 2018). It emphasizes a more open, inclusive, and collaborative research process and encourages new ways of diffusing knowledge by using digital technology. The term "Open Science" often serves as an umbrella term encompassing scholarly outputs, practices, and collaborative digital tools. In its broad understanding, it includes open data, open publications, open educational resources (OER), open source software, open peer review, and citizen science (Bezjak *et al.* 2018). Fecher and Friesike (2014) note the diversity and even ambiguity of the discourse on Open Science and identify several perspectives or "schools of thoughts," ranging from making knowledge freely available for everyone to developing an alternative system for evaluating quality and measuring impact.

Vicente-Sáez and Martínez-Fuentes (2018) acknowledge the diversity of perspectives and concepts of Open Science in their systematic review of the scholarly literature. The authors provide an integrated definition to stimulate a debate about the social, economic, and human added value of Open Science. As a result of their analysis, Open Science is defined as

> the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods. In a nutshell, Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks (Vicente-Sáez & Martínez-Fuentes 2018).

The concept of Open Science and the FAIR data principles have been embraced by the European Commission and incorporated into the European Open Science Cloud roadmap (European Commission 2018). A recent report examines the range of data skills needed to support the implementation of FAIR principles and distinguishes between research community skills, data science, and data stewardship (Hodson *et al.* 2018). The proponents of Open Data recognize that not all data can be open and acknowledge the need to balance openness and protection of sensitive data (European Commission 2016). Qualitative and personal data in social and health sciences pose many challenges for sharing. Some data can be anonymized and released while other data sets will need to remain closed. The European Commission promotes the principle that data should be "as open as possible, as closed as necessary" (European Commission 2016, p.4). Research data management is a critical component of opening and sharing data and determining the levels of openness.

Academic libraries and RDM

The data-intensive research environment and the movement towards Open Science present new opportunities for library profes-

sionals. University libraries in many countries have been assuming leadership roles in promoting open access and offering services in RDM. Traditionally, libraries provided data services for their users by acquiring datasets and ensuring their discovery and access. The new environment challenges libraries to move beyond the traditional service roles of facilitating the discovery and delivery of information resources (Fearon *et al.* 2013). It encourages a more participatory role in the research process and the development of new services to actively support scholars in managing and preserving research data. The concept of data life-cycle plays a central role in developing and organizing RDM consultative and technical services (Carlson 2014). Librarians offer unique expertise in metadata and archiving, and add value at different points of the data cycle.

Academic libraries began to provide a broader range of data management services to support researchers in meeting the requirements of funders and publishers in the last decade. Academic librarians with expertise in RDM who support researchers in meeting funders' compliance and preparing data for release are a vital part of the services. The development of RDM services and the roles of academic libraries in data stewardship have been the subject of extensive survey research (Cox & Pinfield 2014; Tenopir, Birch, & Allard 2012; Tenopir *et al.* 2015). The focus of this research was on the types of services offered by academic librarians, maturity levels, and plans for future development. The findings indicate that academic libraries mostly offer consultative services and training, especially for data management planning. Technical services that involve maintaining a data repository and support for data archiving were limited. Many researchers see RDM services as an extension of traditional academic library roles in outreach and training.

Most of the research, however, focused on academic libraries in the United States and the United Kingdom. More recently, Tenopir *et al.* (2017) conducted a survey of research data services in European academic libraries. The study indicates that more European libraries currently offer consultative than technical services, but also manage infrastructure for data storage and collaborate with other units on campus. Cox *et al.* (2017) expanded the coverage

to seven countries and provided an international comparison of several aspects of RDM development, including policy and governance, type of services, and staff deployment and skills. The IFLA Data Curation project built upon this prior research and expanded it by providing an international and interdisciplinary perspective. The design of the study and the findings are reported in the forthcoming paper (Tammaro *et al.* forthcoming). The preliminary findings about the types and structure of RDM services were presented at the Association for Information Science and Technology conference (Matusiak & Sposito 2017).

IFA DATA CURATION PROJECT

The primary objective of the IFLA LTR project was to identify the roles and responsibilities of RDM practitioners working in multiple countries. The study also focused on the terminology used to describe the emerging practices and new professional roles. The study was designed using a mixed-method approach and consisted of three phases:

- Comprehensive literature review and data mining to analyze the terminology used to describe the emerging practices and new professional roles
- Quantitative content analysis of job announcements for data curators and RDM librarians
- Semi-structured interviews with professionals working as data librarians, data curators, or research data managers.

The quantitative phase of the study concentrated on the content analysis of job announcements derived from a variety of library and information science job posting sites, including International Association for Social Science Information Services and Technology (IASSIST), and Code4Lib. The goal of the content analysis was to examine the titles, roles, responsibilities, qualifications, and competencies listed in the advertised positions. The data set included 441

job advertisements. Most of the analyzed positions (73.6%) were based in the United States. However, the data set also had some international coverage. The widest distribution came from Europe with 17 European countries in the sample.

The findings from the quantitative analysis of job announcements indicate a wide variation in titles used to identify positions. There was no single title standing out as a standard for the discipline. The most common titles included librarianship in some form, such as Data Services Librarians, Digital Scholarship Librarians, or Research Data Management Librarians. The positions were frequently advertised under a wide variety of titles often with additional data-related responsibilities, such as data science or data reference services. In the analyzed data set, RDM services were located primarily (84.2%) in universities and academic libraries. The range of responsibilities also reflects the influence of librarianship with the top responsibilities in public services including instruction, reference, and outreach. However, a degree in librarianship was required in only 27% of the job advertisements.

In the qualitative phase, semi-structured interviews were conducted with professionals working as data librarians, data experts, data curators, or research data managers. The goal of interviews was to gain insight into the practice of research data management and to examine the services from the perspective of the professionals working in the field. The interviews were conducted with 26 professionals from Australia, Canada, U.S. and six countries in Western Europe. The study participants were employed at 24 organizations, including:

- Academic libraries (19)
- Campus-wide research data service centers (3)
- University departments (2)
- Data archive (1)
- Research center (1).

All participants held Masters degrees, including 15 had Masters in Library and Information Science (MLIS). Ten participants had PhDs in a variety of disciplines, including biology, environmental

science, history, information science, medical informatics, or philosophy. The participants held different position titles although many of their responsibilities and job functions overlapped. Several participants, working mostly in Europe, did not have MLIS but had advanced disciplinary degrees and prior research experience. The variety of titles confirmed the findings from the quantitative phase of the study.

Despite the differences in position titles and terminology, the study found a sense of a shared purpose or even mission among the participants. The professionals across institutional and national settings emphasized that their primary roles and responsibilities involved assisting researchers in meeting funder requirements, improving data management practices, and ultimately contributing to a more efficient research process and better-quality data. Several participants mentioned the end-goal of "making data more usable" (P-L, Interview), and efforts to advocate the FAIR data principles. The participants emphasized that although assisting researchers with meeting funder's requirements was one of the immediate goals, they also wanted to improve research practices, as stated by Participant V, "that's really what we want to be leading to, it's not just about compliance but actually trying to change research culture and get people to think it's good research practice" (P-V, Interview).

The types of RDM services identified in this study encompassed both consultative and technical services. The concept of the research data lifecycle played a central role in organizing and structuring services. All professionals participating in this study were engaged in consultative services, outreach, and open access advocacy. The consultative, informational services were typically offered at the beginning of the research cycle in the form of one-on-one consultations, workshops and seminars for faculty and graduate students, or online tutorials and guidelines. The consultative services focused on offering guidance and support in:

- Meeting compliance with funders' requirements
- Developing data management plans (DMP)
- Following data management best practices
- Adhering to data citation standards
- Promoting open access and data sharing

A smaller number of participants assisted researchers with technical aspects of depositing data in repositories and archival storage. Technical services were usually offered at the end of the research data life cycle. Technical infrastructure and the level of support depended on institutional settings. Technical services involved offering support in:

- Data management
- Data formats and file naming conventions
- Data cleaning and verification
- Data conversion
- Data description and documentation
- Metadata creation using standardized schemas
- Data deposit/publishing
- Ingest into repository systems
- Assigning identifiers
- Data anonymization
- Data security
- Archiving and preservation

The participating information professionals often acted as mediators between different stakeholders building networks of expertise and community around good research practices. Their work required some technical skills and knowledge of new technological solutions since they often made recommendations to researchers and led RDM initiatives at their institutions. The new and evolving character of the positions required expertise in multiple areas and the ability to adapt to the changing environment. Specific technical expertise and the level of required skills depended on institutional settings. The study participants emphasized that it's

often impossible for one person to fulfill all the necessary skills and competences found in job descriptions. The lack of technical skills and hands-on experience with databases and scripting was mentioned for professionals with library backgrounds.

RDM services were primarily located in academic libraries as part of research and consultation departments or digital scholarship units. University libraries represented that largest group in the sample but the type of services, the stage of its development, and the level of support for researchers varied greatly between the sites. In the early stage of RDM development, academic libraries usually focused on needs assessment, outreach, training, and open access advocacy and provided consulting services on developing DMPs, metadata, and data curation practices. Academic libraries with more advanced RDM services offered not only assistance with DMPs, metadata, but also with data citation, data sharing and with technical aspects of depositing data in repositories.

The study, however, demonstrated that academic libraries are not the only centers of RDM services on university campuses. It identified new organizational strategies, including embedded services, distributed networks of RDM expertise, and multi-purpose research data services centers. In the embedded model, librarians were working on the faculty-led research projects and research labs throughout the university. They provided support not only at the beginning and end of the research cycle, but also shared expertise and advice on best data management practices throughout the research projects. Distributed networks often had formal structures and were comprised of professionals with expertise in RDM, IT, copyright, research ethics, and scholarly communication. Academic librarians often served as coordinators and referred researchers to the relevant "pockets of expertise" in the campus network. Distributed networks represented efforts in community building around improving data management practices and opening data.

Campus-wide research data service centers represent a new model that reflects an evolution of services and recognition that a more comprehensive suite of skills and expertise is necessary to support data management. Three cases were identified in the sam-

ple – one in the United States and two in Europe. Both European data service centers have evolved from RDM services originally located at academic libraries. These new interdisciplinary initiatives involved cross-campus collaboration and cooperation of several units, including the university library, IT department, legal services, and office for research. Research data service centers tended to be multi-purpose and provided university research communities not only with the expertise, tools, and infrastructure necessary to manage research data but also offered support for other forms of scholarly activities. Academic librarians were employed there along IT specialists and legal experts.

The findings of the study indicate that RDM is an evolving sociotechnical practice that involves not only technical systems and services structured around research data life cycle but also a range of social activities. The work of RDM professionals in improving data management practices and advocating open access occurs on multiple levels, starting with individual researchers and their teams, building networks at their institutions, and then expanding to regional, national, and international communities. The theme of shared values and changing research culture was discussed by participants from multiple countries, pointing to the emerging international character of the RDM profession. Community building emerged as an essential requirement for research data management and involved a shared understanding of the benefits of managed data and the impact of open data on scholarship and society.

Conclusion: opportunities and challenges for the library field

The role of academic libraries in leading and developing RDM services emerged as an important theme in the IFLA Data Curation project and in prior research (Cox & Pinfield 2014; Cox *et al.* 2017; Tenopir *et al.* 2015; 2017). The library and information science (LIS) field can take advantage of the demand for information professionals with knowledge of the research process and skills in ma-

naging and curating data. The report prepared for the European Open Science Cloud points to a shortage of data experts, estimating that half a million specialists with expertise in managing data will be needed to support researchers in the European Union (Ayris *et al.* 2016). The new data-intensive research environment and the global Open Science movement offer opportunities to expand library services beyond the traditional service roles in reference and instruction. Librarians can actively participate in the research process and contribute their unique expertise in information organization, metadata, and archiving. RDM services can also utilize library experience in outreach, open access advocacy, and training.

RDM also poses a set of new challenges for libraries as the field is still in an emergent phase. The development of RDM services at academic libraries involves restructuring and substantial investment in staff and resources. It requires building technical infrastructure for data storage and publishing and forming collaborative partnerships with multiple stakeholders on campus. The model of academic libraries serving as a center of RDM services is prevalent but not the only one. As the findings of the IFLA Data Curation project indicate, the organizational models have been evolving and new approaches are emerging where librarians are embedded in research projects or are becoming partners in campus-wide networks or research data services centers. The new models require strong collaborative skills and building bridges between a library, information technology unit, legal services, and other departments on campus.

The roles, responsibilities, and competencies of RDM librarians are not clearly defined and the practices continue evolving. RDM requires diverse expertise, not only in metadata and information organization standards but also technical skills. RDM creates a demand for information professionals with skills in managing and curating data and with an understanding of the scientific process and research methods. The findings of the IFLA Data Curation project point to some competency gaps in the traditional LIS education, especially in technical training and research methods. RDM as a new area of responsibility for librarians and information professionals requires a combination of technical, instruction, re-

search, and digital archiving skills. Academic librarians have expertise in many areas but also need to acquire new skills and knowledge through expanded professional development. LIS education could also respond to the demand for data experts by developing new programs and concentrations in RDM.

References

Ayris, Paul, Jean-Yves Berthou, Rachel Bruce, Stefanie Lindstaedt, Anna Monreale, Barend Mons, Yasuhiro Murayama, Caj Södergård, Klaus Tochtermann, and Ross Wilkinson. "Realising the European Open Science Cloud." The Commission High Level Expert Group on the European Open Science Cloud, 2016. Accessed November 3, 2018. https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Bezjak, Sonja., April Clyburne-Sherin, Philipp Conzett, Pedro L. Fernandes, Edit Görögh, Kerstin Helbig, Bianca Kramer, and Lambert Heller. "Open Science Training Handbook (Version 1.0)." (2018). Accessed November 5, 2018. https://open-science-training-handbook.gitbook.io/book/#how-to-refer-to-the-handbook.

Borgman, Christine L. "The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059-1078.

Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World.* Cambridge, MA: MIT Press, 2015.

Bornmann, Lutz, and Rüdiger Mutz. "Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References." *Journal of the Association for Information Science and Technology* 66, no. 11 (2015): 2215-2222.

Carlson, Jake. "The Use of Life Cycle Models in Developing and Supporting Data Services." In J. M. Ray (Ed.), *Research Data Management. Practical Strategies for Information Professionals.* West Lafayette: Purdue University Press, 2014., 63-86.

Cox, Andrew M., and Stephen Pinfield. "Research Data Management and Libraries: Current Activities and Future Priorities." *Journal of Librarianship and Information Science* 46, no. 4 (2014): 299-316.

Cox, Andrew M., Mary Anne Kennan, Liz Lyon, and Stephen Pinfield. "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* 68, no. 9 (2017): 2182-2200.

European Commission. "Implementation Roadmap for European Open Science Cloud." (2018). Accessed November 10, 2018. https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf.

European Commission. "Guidelines on FAIR Data Management in Horizon 2020." (2016). Accessed November 10, 2018. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

European Research Council (ERC). "Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in Projects Supported by the European Research Council under Horizon 2020." (2017). Accessed November 3, 2018. https://erc.europa.eu/sites/default/files/document/file/ERC%20Open%20Access%20guidelines-Version%201.1._10.04.2017.pdf.

Fearon, David, Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans. "ARL Spec Kit 334: Research Data Management Services." Washington, DC, Association of Research Libraries, 2013.

Fecher, Benedikt, and Sascha Friesike. "Open Science: One Term, Five Schools of Thought." In: Bartling S., Friesike S. (eds) *Opening Science.* Springer, Cham, 2014.

FOSTER. "Open Science." (2018). Accessed November 5, 2018. https://www.fosteropenscience.eu/taxonomy/term/7.

Hodson, Simon, Sandra Collins, Françoise Genova, Natalie Harrower, Sarah Jones, *et al.* "Turning FAIR Data into Reality." Interim Report of the European Commission Expert Group on

FAIR Data, 2018. Accessed November 10, 2018. https://doi.org/10.5281/zenodo.1285272.

Kraker, Peter, Derick Leony, Wolfgang Reinhardt, and Günter Beham. "The Case for an Open Science in Technology Enhanced Learning." *International Journal of Technology Enhanced Learning* 3, no. 6 (2011): 643-654.

Matusiak, Krystyna. K. and Frank Andreas Sposito. "Types of Research Data Management Services: An International Perspective." *Proceedings of the Association for Information Science and Technology* 54, no. 1 (2017): 754-756.

Molloy, Jennifer C. "The Open Knowledge Foundation: Open Data Means Better Science." *PLoS Biology* 9, no. 12 (2011): e1001195. Accessed November 10, 2018. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001195.

National Science Foundation (NSF). "Dissemination and Sharing of Research Results." Accessed November 3, 2018. https://www.nsf.gov/bfa/dias/policy/dmp.jsp.

Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck *et al.* "Promoting an Open Research Culture." *Science* 348, no. 6242 (2015): 1422-1425.

Pryor, Graham, Sarah Jones, and Angus Whyte, eds. *Delivering Research Data Management Services: Fundamentals of Good Practice.* London, Facet Publishing, 2013.

Swan, Alma. "Policy Guidelines for the Development and Promotion of Open Access." UNESCO, 2012. Accessed November 4, 2018. http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/policy-guidelines-for-the-development-and-promotion-of-open-access/.

Tammaro, Anna Maria, Krystyna K. Matusiak, Frank Andreas Sposito, and Vittore Casarosa. "Data Curator's Roles and Responsibilities: An International Perspective." *Libri* (forthcoming).

Tenopir, Carol, Ben Birch, and Suzie Allard. "Academic Libraries and Research Data Services. Current Practices and Plans for the Future." An ACRL White Paper Chicago: Association of College and Research Libraries (2012). Accessed November 5, 2018. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf.

Tenopir, Carol, Dane Hughes, Suzie Allard, Mike Frame, Ben Birch, Lynn Baird, Robert Sandusky, Madison Langseth, and Andrew Lundee. "Research Data Services in Academic Libraries: Data Intensive Roles for the Future?." *Journal of eScience Librarianship* 4, no. 2 (2015): 4.

Tenopir, Carol, Sanna Talja, Wolfram Horstmann, Elina Late, Dane Hughes, Danielle Pollock, Birgit Schmidt, Lynn Baird, Robert J. Sandusky, and Suzie Allard. "Research Data Services in European Academic Research Libraries." *Liber Quarterly* 27, no. 1 (2017): 23-44.

UK Research and Innovation. "Common Principles on Data Policy." Accessed November 3, 2018. https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/.

Vicente-Saez, Ruben, and Clara Martinez-Fuentes. "Open Science now: A Systematic Literature Review for an Integrated Definition." *Journal of Business Research* 88 (2018): 428-436.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg *et al.* "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (2016).

Integración de los principios de *linked data* en el registro bibliográfico

Eder Ávila Barrientos
Universidad Nacional Autónoma de México

Introducción

L*inked Data* es un conjunto de buenas prácticas para publicar y vincular datos estructurados en el entorno de la web.

> *Linked Data* extiende los principios de la *World Wide Web* desde la vinculación de los documentos hasta la de vincular piezas de datos y crear una Web de Datos; especifica los datos y sus respectivas relaciones, y le proporciona datos procesables por máquina a Internet. Está basado en Técnicas estándar web, pero las amplía para proporcionar el intercambio de datos y la integración. (Sakr, Sherif, *et al.* 2018, 9).

El informe sobre datos bibliotecarios enlazados del Grupo incubadora del W3C *(Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets),* es uno de los desarrollos más significativos que explica la interacción entre los principios de *Linked Data* en el ambiente de las bibliotecas, y se trata de un documento relevante para entender la oportunidad que se abre en las bibliotecas para alcanzar el impacto digital que

pretenden, además de concretar nuevos modelos en el acceso y la organización de los recursos de información.

Según Issac (2011), el Library Linked Data Incubator Group tiene como misión analizar la situación de los modelos y esquemas de metadatos y los estándares y protocolos de interoperabilidad que se deberían usar para la publicación y el uso de *Linked Data* con datos de bibliotecas.

Los subgrupos creados como parte del Library Linked Data Incubator Group, quedaron divididos de la siguiente manera:

Primer grupo. Encargado del tratamiento de datos bibliográficos de las Bibliotecas Nacionales, como la Biblioteca Británica, y las de Francia, Alemania y España.

Segundo grupo: Encargado de tratar los datos de autoridades de las mismas bibliotecas.

Tercer grupo. Encargado de conformar los diversos vocabularios controlados, la mayoría de ellos relativos a ontologías.

Cuarto grupo. Se enfoca al estudio de la presencia de recursos de información en la nube de datos enlazados (la cual se representa mediante el Linked Data Cloud Diagram).

Quinto grupo. Encargado del estudio de la relación de las citas en artículos científicos. Este estudio otorga la posibilidad de construir conjuntos de datos científicos y de mostrar sus respectivas relaciones a partir del análisis de las citas para cada artículo. Esta labor es muy interesante, ya que permite concebir nuevos e innovadores servicios de información digital y servicios documentales que apoyen el proceso de investigación, experimentación y análisis de los investigadores en diversas áreas del conocimiento.

Es importante mencionar que muchas publicaciones científicas que tienen presencia en el entorno digital utilizan metadatos para describir los datos de su publicación, lo cual facilita la creación de relaciones y en consecuencia el enlace entre datos y la creación de conjuntos o datasets.

Sexto grupo. Este grupo se encarga de darles tratamiento a los objetos digitales disponibles en el contexto de las bibliotecas. El objetivo del grupo consiste en la creación de un repositorio digital que vincule los objetos a partir del uso de metadatos semánticos.

Séptimo grupo. Se enfoca a la investigación sobre la construcción de colecciones digitales. En este grupo se abordan los alcances que FRBR puede proporcionar para la descripción de colecciones digitales. FRBR es un modelo conceptual que, al momento de ser aplicado en la organización de recursos de información, es susceptible de convertirse en un modelo descriptivo con diferentes niveles de descripción.

Octavo grupo. Aborda la presencia de la biblioteca en el entorno de las redes sociales. Un entorno muy atractivo que permite el intercambio de información entre pares y en donde los datos enlazados pueden tener una función significativa en la satisfacción de las necesidades de información de los usuarios remotos de las bibliotecas.

En suma, el trabajo de los ocho grupos se traduce en tres principales áreas de investigación:

- Área de preparación de los datos. Que se enfoca a la creación de herramientas que permitan transformar, almacenar y vincular los datos de las bibliotecas.
- Área de definición de normas. Que se encarga de abordar la construcción de normas que permitan controlar y uniformar el proceso de creación de datos enlazados.
- Área de desarrollo de interfaces. Encargada del diseño de interfaces de búsqueda y recuperación de información. La interfaz de cualquier sistema de información digital es de suma relevancia para que el usuario remoto pueda tener acceso a la información en el entorno digital.

De esta manera, se requiere contar con un método de recuperación de información que permita identificar, descubrir y acceder a los datos documentales que están disponibles en el ambiente web. Los datos de las bibliotecas que están representados en los registros bibliográficos remiten a recursos de información que pueden tener patrones de vinculación entre los datos documentales que están disponibles en diversas fuentes de la web.

Se estima que la aplicación de los principios de *Linked Data* puede favorecer el desarrollo de un método para la recuperación

de información en este ambiente. Por lo tanto, es preciso responder a tres preguntas elementales:

> ¿Cómo se aplican los principios de *Linked Data* en el registro bibliográfico?
> ¿Cómo se vinculan los datos disponibles en estos registros con otras fuentes de datos disponibles en la web?
> ¿Qué alcances y limitaciones tiene esta vinculación?

El propósito de este trabajo consiste en analizar la integración de los principios de *Linked Data* en el registro bibliográfico para identificar patrones de vinculación entre los datos disponibles en las bibliotecas y el ambiente web.

Revisión de la literatura

La integración de los principios de *Linked Data* en el registro bibliográfico es un tema que ha sido abordado con anterioridad. A través de la formulación teórica de la web semántica y su relación con las bibliotecas, pueden localizarse hallazgos significativos que intentan explicar dicha integración mediante estudios de caso y reflexiones teóricas, que exponen la evolución de los principios de la organización de la información y la fuerte influencia que ejercen los principios de la web semántica sobre estos postulados.

Por ejemplo, Alemu *et al.* (2012) realizaron un análisis teórico que sugiere recomendaciones para llevar a cabo un cambio conceptual de los metadatos centrados en el documento a metadatos centrados en los datos. A su vez, discutieron la importancia de ajustar los modelos de las biblioteca actuales, como RDA y FRBR a los modelos basados en los principios de los datos enlazados.

Los datos bibliográficos requieren ser tratados de manera independiente, pero inherente al recurso al cual pertenecen. Pues de esta manera el análisis de los datos bibliográficos permitirá establecer un vínculo significativo entre aquellos datos que contengan atributos similares. Para ello, es necesario que los principios bi-

bliotecológicos para la organización de la información sean compatibles con los principios de *Linked Data*.

Cole *et al.* (2013, 189) "[…] evaluaron la viabilidad y los desafíos de transformar los registros bibliográficos de bibliotecas tradicionales en *Linked Data*". Como parte de este estudio, los autores identificaron la falta de compatibilidad entre MARC y los principios de los datos enlazados. No obstante, esta incompatibilidad ha tratado de erradicarse mediante la adopción de estrategias que tienen el propósito de integrar ambos principios. Por ejemplo, en la generación de vocabularios para la representación de los datos.

Por su parte, Tillet (2013,140) manifiesta que

> […] los datos de la biblioteca sobre nuestros recursos ya no sólo deben almacenarse en cajones de catálogo como herramienta de inventario para acceder a las colecciones de una biblioteca individual. Ahora se puede poner a disposición de cualquier persona, en cualquier lugar del mundo, en cualquier momento.

En este sentido, la disponibilidad de los datos de la biblioteca en el ambiente web permite la posibilidad de vincularlos con otras fuentes disponibles en este contexto. Siempre y cuando los registros bibliográficos se adapten al entorno actual de la web. Pues "[…] los datos enlazados ofrecen la posibilidad de realizar una reestructuración profunda del registro bibliográfico que se presenta con una nueva estructura granular" (Iacono 2014, 80).

La granularidad de los datos es una característica que les permite incrementar el nivel de detalle en su descripción y estructuración. En consecuencia, la aplicación de los principios de *Linked Data* en el registro bibliográfico beneficiará la función del registro como un medio para establecer la vinculación semántica de los datos mediante una arquitectura interoperable para la interconexión de fuentes de datos disponibles en la web.

A su vez, se ha identificado la función de *Linked Data* como método para la óptima recuperación de información en las bibliotecas. Mitchell (2016) descubrió que en los últimos dos años se

habían realizado importantes investigaciones y publicaciones que documentaban proyectos técnicos específicos, aplicaciones, vocabularios y mejores prácticas de la comunidad bibliotecaria en relación con los datos enlazados y su interacción en las bibliotecas.

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. Pero los procesos de búsqueda y recuperación de la información han evolucionado, actualmente se requiere de métodos integrales que permitan el descubrimiento de información en diversos contextos relacionados.

Por lo tanto, la integración de *Linked Data* en el registro bibliográfico pone de manifiesto la interacción de normas para la descripción de los recursos, formatos de codificación y principios semánticos. RDA es la norma de descripción que en un futuro no muy lejano será aplicada en su totalidad en el ambiente de las bibliotecas. El formato MARC es el esquema de codificación por excelencia que es utilizado en los sistemas integrales de gestión de bibliotecas y es un elemento importante para la búsqueda y recuperación de recursos de información documental en las bibliotecas. La integración de RDA y MARC con los principios de *Linked Data*, ha sido motivo de análisis y discusión dentro de la literatura especializada.

Faith y Chrzanowski (2015, 133) realizaron un prototipo básico de la aplicación de RDA con RDF. Mediante los resultados obtenidos, los autores manifiestan que "[…] la vinculación de datos para una búsqueda más relevante y habilitada ayuda a abrir las bibliotecas a un mundo más amplio de posibilidades conectadas". Además, los usuarios pueden recibir una navegación más diversa y opciones de búsqueda más sólidas a través de los datos vinculados. Los datos vinculados son un medio para conectar a más personas con información más relevante.

Por su parte, Shieh (2018) informó que el Programa de Catalogación Cooperativa de la Biblioteca del Congreso (siglas en inglés PCC-LC) ha comenzado con el mapeo de elementos entre BIBFRAME,

RDA y MARC con el propósito de mejorar de las prácticas correctas para los sistemas de recuperación de información emergentes, así como los estándares actuales de descripción actuales.

Los resultados de la búsqueda y recuperación de literatura que aborda la aplicación de los principios de *Linked Data* en el registro bibliográfico son considerables en cantidad. Sin embargo, al momento de analizar con detalle los trabajos publicados, se observa una laguna teórica y pragmática relacionada con la aplicación formal de los principios en dicho registro. La revisión de la literatura del objeto de investigación planteado en este trabajo ha permitido identificar que las breves reflexiones teóricas del objeto de estudio y las mínimas pruebas de aplicación hacen más complejo el estudio de la temática planteada. Por lo tanto, se requiere de mayor investigación teórica y metodológica para descubrir los patrones de comportamiento que se generan mediante la aplicación de los datos enlazados en el registro bibliográfico.

Manejo de datos enlazados en las bibliotecas

La revolución de los datos es un fenómeno que ha sido provocado por el impacto de las tecnologías digitales en contextos donde la información es un elemento trascendental para la generación de nuevos conocimientos. En la actualidad, los datos son producidos a gran velocidad y de manera continua por personas, computadoras y como parte del uso de aplicaciones comerciales y de geolocalización.

Smith (2014) ha explicado con anterioridad la utilización de métodos y tecnologías digitales para el procesamiento de los datos disponibles en las bibliotecas. Uno de los aspectos mas sobresalientes del procesamiento de los datos, recae en la cercana relación que tiene con los procesos de descripción y catalogación de los recursos.

Se puede afirmar que la catalogación y descripción de los recursos son métodos analíticos e intelectuales para el procesamiento y la obtención de datos bibliográficos. Las bibliotecas contienen enormes cantidades de datos de índole bibliográfica y documental que son registrados con formatos altamente especializados. Los

datos de la biblioteca son generados como parte de procesos intelectuales mediante el uso de normas, vocabularios y principios de índole bibliotecológica.

Los datos de la biblioteca son utilizados como puntos de acceso para los recursos de información documental que son registrados en los catálogos. Estos datos son almacenados y registrados en herramientas para la búsqueda y recuperación de información. Por lo tanto, los registros bibliográficos de las bibliotecas se convierten en una fuente para la construcción de datos enlazados en el ambiente de las bibliotecas.

El manejo de datos enlazados pone de manifiesto un método para la eficiente publicación y vinculación de datos estructurados en la web. Los datos que pertenecen a los registros bibliográficos de las bibliotecas son de gran utilidad para la generación de datos enlazados de tipo bibliográfico y de autoridad.

Los registros bibliográficos codificados en formato MARC deben adaptarse a los principios de *Linked Data*. el proceso de adaptación de estos registros es sistematizado y se encuentra relacionado con el uso de normas y estándares de índole internacional. La desfragmentación del registro bibliográfico dará como resultado el tratamiento individual de los datos que caracterizan a los recursos de información documental.

Cada dato bibliográfico puede vincularse semánticamente con otros datos de atributos similares y que estén disponibles en el ambiente web. Para llevar a cabo esta vinculación es necesario ejercer buenas prácticas de manejo de datos enlazados en las bibliotecas. El proceso para el eficiente manejo de datos enlazados se compone de las siguientes fases:

A. Fase de selección de datos. Los registros de datos bibliográficos que serán seleccionados pueden pertenecer a una determinada colección. Formar parte de una temática en un dominio específico de conocimientos. La especificidad de los datos seleccionados es un asunto importante, pues a mayor especificidad en los datos será mayor el grado de exactitud que se alcance al momento de vincularlos.

Una vinculación semántica de datos es un procedimiento intelectual que además de conectar a los datos, tiene el propósito de explicar el significado de la relación que se establece entre ellos.

B. Fase de normalización de los datos. La estructuración de los datos es un procedimiento normalizado. Para ello se utilizan los principios de *Linked Data* señalados por el W3C. La asignación de URIs a los datos, la codificación de los datos con RDF y la utilización de vocabularios estandarizados son algunas de las acciones que se desarrollan en esta fase. Además, se debe contemplar el uso de los principios bibliotecológicos y los lenguajes documentales que favorezcan la estructuración de los datos y beneficien su descripción y acceso.

C. Fase de descripción de los datos. El registro de los datos debe ser exacto, sin ambigüedades y libre de inexactitudes. RDA y Dublin Core, proporcionan elementos descriptivos para representar los atributos bibliográficos y de contenido de los recursos. Los tesauros, folksonomías y ontologías pueden favorecer la descripción temática de los recursos. La óptima descripción temática del recurso permitirá obtener datos de mayor precisión para establecer vinculaciones de mayor significado entre los datos.

D. Fase de vinculación de los datos. Los datos debidamente descritos y estructurados deberán vincularse entre sí. Las vinculaciones de los datos deben explicar la relación que existe entre ellos y el significado que los rodea en un determinado contexto. Pues no se trata únicamente de establecer conexión entre los datos, sino de explicar el significado de la vinculación que se establece entre los datos disponibles en diversas fuentes. Para vincular los datos es necesario establecer interoperabilidad entre las fuentes que serán conectadas.

E. Fase de recuperación y acceso a los datos. Los datos enlazados generados en las bibliotecas deben ser abiertos, libres de cualquier restricción técnica, legal y económica. Para

ello, deberán aplicarse los principios de licencias abiertas de datos. La recuperación de los datos enlazados deberá permitir el descubrimiento de los datos y de sus respectivas vinculaciones. Además de la tradicional búsqueda textual, será necesario visualizar gráficamente las vinculaciones mediante el uso de una interfaz para la consulta de grafos.

F. Fase de preservación de los datos. La prospectiva del uso de los datos debe sujetarse a un proceso planificado. Se debe contemplar qué datos será necesario conservar para ser utilizados en un futuro y garantizar su acceso sin que importen los rápidos cambios tecnológicos del contexto que los rodea. Es deseable contar con una política de preservación de datos que respalde el proyecto de datos enlazados y su generación en las bibliotecas.

El óptimo manejo de los datos enlazados en las bibliotecas pone de manifiesto la figura de un cambio de paradigma relacionado con la catalogación de los recursos de información. Pues se estima que la aplicación de *Linked Data* en el ámbito de la organización de la información, da la pauta para la descripción semántica de los recursos; en el siguiente apartado se abordan algunos de los principios identificados que explican la formulación de este proceso.

Hacia la descripción semántica de los recursos

RDA (2014) menciona tres tipos de descripciones de recursos:

A. Descripción compresiva. Se utiliza para describir a los recursos como un todo. Se emplea para describir cualquier tipo de recurso.
B. Descripción analítica. Es utilizada para describir una parte de un recurso más amplio.
C. Descripción jerárquica. Une los dos tipos anteriores, es decir, combina una descripción integral de un recurso, con la descripción analítica de una o más de sus partes, por ejem-

plo: se describe un libro y además un capítulo del mismo; una revista científica y sus respectivos artículos.

La descripción semántica del recurso reúne los tipos de descripciones anteriores y explica la vinculación de los datos que son descritos y que pertenecen a los registros. Pues no se trata únicamente de establecer relaciones superficiales entre los recursos. Sino de explicar la conexión y darle significado a la vinculación que se establece entre los datos de los recursos.

Por lo tanto, la descripción semántica de los recursos se define como un proceso intelectual y apegado a normas que tiene el propósito de registrar y representar los atributos bibliográficos, temáticos y de autoridad de los recursos para explicar el significado de las vinculaciones existentes entre los datos pertenecientes a estos recursos.

Indudablemente, los métodos y estándares para organizar la información han cambiado, y se han adaptado a las características de los diferentes tipos de recursos que han surgido y han incorporado el uso de las tecnologías digitales para su consulta. Sin embargo, los principios sustanciales en los que descansa la organización y recuperación de la información continúan siendo los mismos. Localizar la información, identificar si es la información que necesitamos para tomar la decisión de obtenerla. En estos principios descansan desde el desarrollo de los catálogos hasta el de la web semántica (Martínez 2009, 12).

Los principios sustanciales de la organización de la información deben evolucionar de acuerdo con las exigencias del contexto de información actual. Hoy en día los datos de las bibliotecas están cobrando mayor relevancia para satisfacer las demandas informativas de la comunidad.

Actualmente los datos de la biblioteca remiten a una amplia gama de recursos de información documental. Algunos de estos recursos se vinculan de manera directa con otros contextos fuera de la biblioteca, por ejemplo, con bases de datos, repositorios y plataformas de contenidos digitales. Es necesario definir métodos que permitan identificar estas conexiones de una manera automatizada y accesible para el usuario final.

Resultados de la integración de Linked Data en el registro bibliográfico

Linked Data reúne los componentes principales para desarrollar la web semántica. Berners-Lee (2006), definió cuatro reglas básicas para construir datos enlazados:

1. Utilizar URIs para nombrar a las cosas disponibles en la web.
2. Utilizar el protocolo HTTP-URI para que los usuarios de la web puedan buscar esas cosas.
3. Cuando un usuario busca un URI, debe proporcionar información útil empleando los estándares RDF y SPARQL.
4. Incluir enlaces a otros URIs para que el usuario pueda descubrir más cosas.

Linked Data es un concepto de propósito general. Literalmente cualquier cosa puede ser descrita utilizando datos enlazados.

> RDF proporciona un modelo común para datos enlazados y es particularmente adecuado para representar datos en la Web. *Linked Data* utiliza RDF como su modelo de datos y lo representa en una de varias sintaxis (Wood, Zaidman & Luke 2014, 9).

Candela, *et al.* (2015) llevaron a cabo un prototipo de implementación de *Linked Data* en un contexto de datos bibliográficos. Migraron 200 mil registros del catálogo de la Biblioteca Miguel de Cervantes a una nueva base de datos relacional cuyo modelo de datos se adhiere a las especificaciones FRBR y FRAD. El contenido de la base de datos se asignó posteriormente a tripletas RDF que emplean el vocabulario de RDA para describir las entidades, así como sus propiedades y relaciones.

A su vez, Possemato (2018), realizó una investigación en donde se ocupa de la aplicación del estándar RDA en el campo de los datos vinculados y de cómo se puede utilizar este estándar para mejorar la calidad de los datos producidos por las bibliotecas y al-

canzar así las ventajas que la web semántica puede aportar al sector del patrimonio cultural.

Las dos investigaciones anteriores, fueron relevantes para comprender el camino a seguir en la implementación de *Linked Data* en el registro bibliográfico. Pues a pesar de haber localizado una considerable muestra de estudios de caso, la mayoría de ellos eran abordados desde una perspectiva informática o enfocada al ámbito computacional y tomaban como objeto de estudio datos de diversa tipología y naturaleza disciplinar.

De esta manera, para integrar los principios de *Linked Data* en el registro bibliográfico, se tomó en cuenta la obra de *El nombre de la rosa,* escrita en el año 1980 por el filósofo italiano Humberto Eco. Se utilizaron los elementos núcleo para la descripción de manifestaciones señaladas en RDA y el formato MARC para obtener dos ejemplificaciones básicas de la construcción del registro.

Se aplicaron los principios básicos de RDF para obtener la estructura general del recurso que permitieran representar a los datos bibliógrafos como un triple. A cada dato del registro bibliográfico le fue asignado un URI. Cada dato bibliográfico debe contar con un URI único e individual que tenga dependencia directa con el sistema que los almacena y genera. En este sentido, cada URI es irrepetible y representa a un dato dentro de un dominio específico. Cada URI debe estar normalizado bajo principios interoperables que le permitan vincularse con otra fuente ajena a la biblioteca. La interoperabilidad entre los datos permitirá consultarlos en diferentes plataformas y dispositivos.

Ilustración 1. Ejemplo de registro MARC con elementos estructurales de *Linked Data*. Fuente: elaboración propia, 2018.

SUJETO URI: http://el_nombre_de_la_rosa_work_rda	
PREDICADO	OBJETO
Elemento MARC	Registro de los datos
http://marc.008/35-37	http://spa
http://marc.020	$a http://ISBN_978-970-810-026-7
http://marc.040	$a http://rda
http://marc.100	$a http://Eco_Umberto $d http://1932-2016 $e http://Autor
http://marc.245	$a http://El_nombre_de_la_rosa $c http://Umberto_Eco
http://marc.250	$a http://Segunda_edición
http://marc.264	$a http://México $b http://Random_House_Mondadori $c http://2004 $c http://1980
http://marc.300	$a http://783_páginas
http://marc.336	$a http://Texto
http://marc.338	$a http://Volumen
http://marc.650	http://Novela_histórica
http://marc.700	$a http://Ricardo_Pochtar $e http://Traductor

Los datos bibliográficos deben ser desfragmentados para luego ser tratados semánticamente bajo los principios de *Linked Data* pero, se deben eliminar puntuaciones, pues éstas son irrelevantes para la búsqueda, recuperación y vinculación de los datos. En el caso del formato MARC, los subcampos e indicadores de codificación, pueden llegar a alterar la representación de los datos en formato RDF. Sin embargo, es posible utilizar los datos que están colocados en las diversas áreas del registro bibliográfico codificados en formato MARC.

Ilustración 2. Ejemplo de registro RDA con elementos estructurales de *Linked Data*. Fuente: elaboración propia, 2018.

SUJETO URI: http://el_nombre_de_la_rosa_work_rda	
PREDICADO	OBJETO
Elemento RDA	Registro de los datos
http://rda.título	http://El_nombre_de_la_rosa
http://rda.mención_de_responsabilidad	http://Umberto_Eco
http://rda.edición	http://Segunda_edición
http://rda.lugar_de_publicación	http://México
http://rda.editor	http://Random_House_Mondadori
http://rda.fecha_de_publicación	http://2004
http://rda.copyright	http://1980
http://rda.identificador_de_la_manifestación	http://ISBN_978-970-810-026-7
http://rda.soporte	http://Volumen
http://rda.extensión	http://783_páginas
http://rda.tipo_de_contenido	http://Texto
http://rda.idioma_de_la_expresión	http://Español
http://rda.creador	http://Eco_Umberto_1932-2016
http://rda.colaborador	http://Ricardo_Pochtar
http://rda.designador_de_relación	http://Traductor
http://rda.relación_temática	http://Novela_histórica

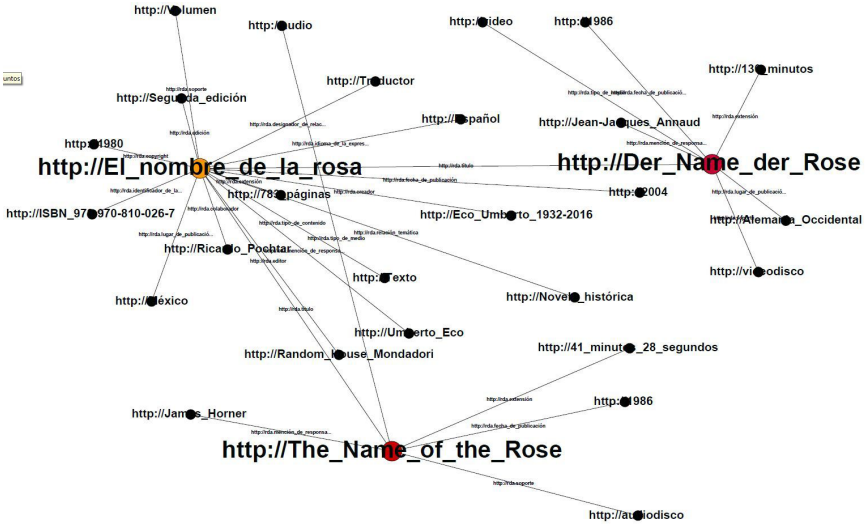
Los elementos de RDA presentan mayor flexibilidad en su integración con los principios de *Linked Data*. Sin embargo, también es necesario omitir la puntuación al momento de registrar los datos. RDA plantea la posibilidad de establecer relaciones bibliográficas entre los recursos. Los designadores de relación son elementos que fomentan el establecimiento de conexiones entre los datos. Sin embargo, se trata de una relación meramente superficial que carece de una explicación semántica.

Cuando los datos del registro bibliográfico son desfragmentados y tratados de manera individual, es posible construir grafos

RDF que ayudan a comprender su comportamiento en un determinado contexto o dominio. Según Powell y Hopkins (2015, 9), "un grafo es un modelo abstracto conceptual que sirve para representar algunos aspectos del mundo que se pueden detectar, observar o inferir." Un grafo puede utilizarse como un método visual para comprender las relaciones que se establecen entre los datos que pertenecen a una realidad específica.

En este sentido, los grafos RDF son desarrollados siguiendo una lógica entre sujeto, predicado y objeto. En la *Ilustración 3,* puede observarse la construcción de un grafo RDF que representa a los datos de los registros RDA y MARC de la obra "El nombre de la rosa". En color naranja, se puede visualizar el nodo correspondiente al título de la obra. El cual se trata de una versión impresa del libro (manifestación RDA). Los nodos rojos representan la videograbación *(Der name der rose)* y grabación sonora *(The name of the rose)* que se desprenden de la obra principal.

Ilustración 3. Grafo de la vinculación de datos bibliográficos disponibles en los registros MARC y RDA. Fuente: elaboración propia, 2018.



Los datos de color azul corresponden a los elementos RDA que tienen la función de predicado en el grafo. Cada elemento RDA explica el significado que tiene la vinculación efectuada entre los datos que están sombreados de color verde. Cada uno de los datos representados en el grafo tiene la capacidad de vincularse con atributos similares. La explicación del significado de la vinculación se expresa en el grafo mediante una visualización integral.

La consulta de los datos enlazados necesita de una visualización grafica que contribuya a la comprensión del comportamiento de los datos que son vinculados en un determinado contexto. En el ámbito de las bibliotecas, los datos bibliográficos se encuentran en constante movimiento, ya sea mediante actualizaciones que sufren o con la generación de nuevos recursos que permiten desarrollar redes de datos bibliográficos más extensas y complejas.

En la actualidad los datos bibliográficos deben contemplarse como elementos que sirven para desarrollar estructuras complejas que pueden vincularse con otros contextos independientes a las bibliotecas. El potencial de los datos bibliográficos dependerá de su capacidad para vincularse en otros contextos pertinentes y significativos que se caractericen por contar con información arbitrada y de calidad. Adicionalmente, los datos deben ser publicados de manera abierta, sin barreras de índole técnica, legal y económica.

Consideraciones finales

La integración de los principios de *Linked Data* en el registro bibliográfico, es un proceso intelectual y fundamentado basado en el uso de normas como RDF, URIs y SPARQL. Se requiere que los datos bibliográficos tengan mayor flexibilidad para adaptarse a entornos interoperables de datos abiertos enlazados. Para ello, es necesario que el registro bibliográfico pueda vincularse con fuentes externas al catálogo de la biblioteca.

Mediante la aplicación de los principios de *Linked Data* en el registro bibliográfico, fue posible identificar la necesaria flexibilidad que requiere RDA y MARC para adaptarse a los principios de

los datos enlazados. Si bien es posible asignar URIs a cada uno de los datos del registro y estructurarlos bajo RDF, es necesario que los subcampos del formato MARC tengan mayor adaptabilidad con los principios de *Linked Data*.

Los elementos de RDA manifiestan una mayor flexibilidad de adaptación a los principios de *Linked Data*. Sin embargo, se requieren ejercicios de integración con mayor grado de complejidad y detalle que permitan identificar el comportamiento de los datos bibliográficos mediante su estructuración semántica. Proyectos como *RDA Registry* han generado vocabularios RDA-RDF que será necesario abordar en futuras investigaciones mediante la generación de prototipos de registro de datos enlazados de índole bibliográfica.

El grafo RDF es un método para visualizar los datos enlazados generados. El acceso y consulta de los datos enlazados pone de manifiesto la generación de este tipo de grafos. Así la consulta de los datos y sus respectivas vinculaciones serán más usables y accesibles mediante este método gráfico.

Se estima que la generación del BIBFRAME por parte de LC y LRM de IFLA, fomenten una mayor adaptabilidad, interoperabilidad y flexibilidad de los datos disponibles en el registro bibliográfico con los principios de *Linked Data*. Sin embargo, será necesario analizar los alcances y limitaciones de ambos desarrollos para conformar entornos de datos enlazados bibliográficos.

Obras consultadas

Alemu, Getaneh, Brett Stevens, Penny Ross, and Jane Chandler. 2012. "Linked Data for Libraries: Benefits of a Conceptual Shift from Library-Specific Record Structures to RDF-Based Data Models." *New Library World* 113 (11): 549–70. DOI:10.1108/03074801211282920.

Berners-Lee, Tim. "Linked data, publicada el 27 de julio de 2006", https://www.w3.org/designissues/linkeddata.html

Candela, G., P., Escobar, M, Marco-Such, R.C., Carrasco. 2015. "Transformation of a library catalogue into RDA linked open data." In *lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics),* 9316: 5–7. DOI:10.1007/978-3-319-24592-8.

Cole, Timothy W., Myung Ja Han, William Fletcher Weathers, and Eric Joyner. 2013. "Library MARC Records Into Linked Open Data: Challenges and Opportunities." *Journal of Library Metadata* 13 (2–3): 163–96. DOI:10.1080/19386389.2013.826074.

Faith, Ashleigh, and Michelle Chrzanowski. 2015. "Connecting RDA and RDF: Linked Data for a Wide World of Connected Possibilities." *Pennsylvania Libraries: Research & Practice* 3 (2): 122–35. DOI:10.5195/PALRAP.2015.106.

Iacono, Antonella. 2014. "Dal Record Al Dato. Linked Data e Ricerca Dell'informazione Nell'OPAC." *Italian Journal of Library, Archives, and Information Science,* 5 (1): 77–102. DOI:10.4403/jlis.it-9095.

ISAAC, Antoine, *et al.* 2011. "Library Linked Data Incubator Group: Datasets, Value, Vocabularies, and Metadata Element Sets: W3C Incubator Group Report", http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

Joint Steering Committee for Development of RDA. 2014. *Resource Description & Access.* American Library Association, Canadian Library Association, CILIP.

Martínez Arellano, Filiberto Felipe. 2009. "Organización de la información: del catálogo a la web semántica". En *Memoria del XXVI Coloquio de Investigación Bibliotecológica y sobre la Información,* 1, 2 y 3 de octubre de 2008 compiladores Filiberto Felipe Martínez Arellano, Juan José Calva González, 3-14. México: UNAM. Centro Universitario de Investigaciones Bibliotecológicas.

Mitchell, Erick. 2016. "Library Linked Data: Early Activity and Development." *Library Technology Reports.* Vol. 52. DOI:10.5860/ltr.52n1.

Possemato, Tiziana. 2018. "How RDA Is Essential in the Reconciliation and Conversion Processes for Quality Linked Data." *Italian Journal of Library, Archives, and Information Science,* 9 (1): 49–61. DOI:10.4403/jlis.it-12447.

Powell, James & Matthew, Hopkins. 2015. *A librarian´s guide to graphs, data and the semantic web.* USA: Elsevier.

Sakr, Sherif, Marcin Wylot, Raghava Mutharaju, Danh Le Phuoc, and Irini Fundulaki. 2018. Linked Data: Storing, Querying, and Reasoning. *Linked Data: Storing, Querying, and Reasoning.* USA: Springer Link. DOI:10.1007/978-3-319-73515-3.

Shieh, Jackie. 2018. "Reports from the Program for Cooperative Cataloging Task Groups on URIs in MARC & BIBFRAME." *Italian Journal of Library, Archives, and Information Science,* 9 (1): 111–20. DOI:10.4403/jlis.it-12429.

Smith, K.M. (2014). *Handbook of data processing for libraries: modern methods and latest technologies.* London: Koros Press.

Tillett, Barbara. 2013. "RDA and the Semantic Web, Linked Data Environment." *Italian Journal of Library & Information Science* 4 (1): 139–45. DOI:10.4403/jlis.it-6303.

Wood, David, Marsha, Zaidman y Ruth, Luke. 2014. *Linked data: structured data on the web.* Estados Unidos de América: Manning.

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:

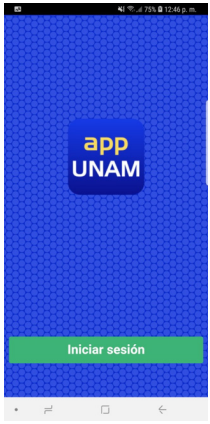
- Rebasan las capacidades de las herramientas disponibles en las áreas académicas y administrativas para su realización.
- Involucran el uso de *software* especializado *(frameworks)* y plataformas de cómputo de alto rendimiento *(clusters),* destinados hoy en día sólo a la investigación científica.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

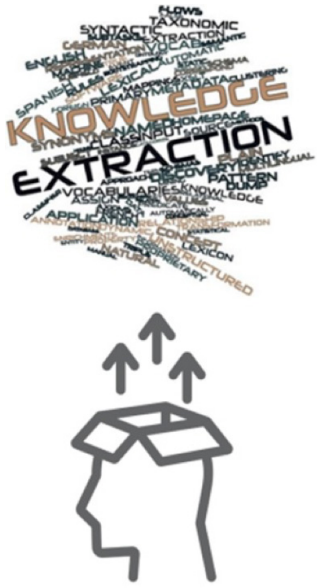
- La producción de datos masivos en la UNAM hoy en día ya es una realidad.
- Que requiere de las tecnologías de Ciencia de Datos y Big Data para su manejo y explotación.

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 autoconsumo (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

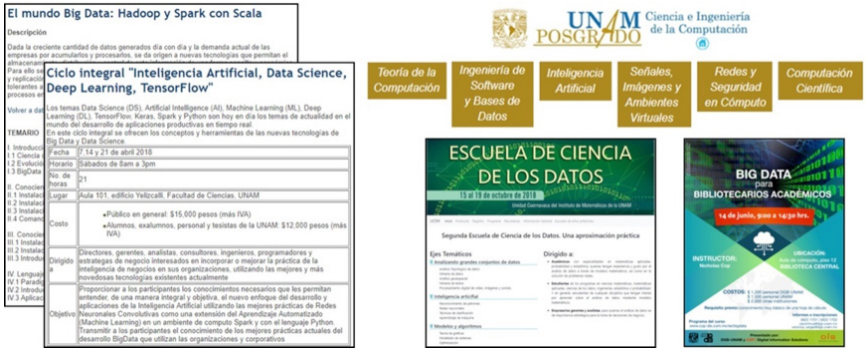
3. En lo referente a la infraestructura disponible para la docencia en Ciencia de Datos y Big Data:

- 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



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.