

Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales

Perla Olivia Rodríguez Reséndiz
Coordinadora



Q335 Inteligencia artificial y datos masivos en archivos digitales
I57 sonoros y audiovisuales / Coordinadora Perla Olivia Rodríguez
Reséndiz. - México: UNAM. Instituto de Investigaciones
Bibliotecológicas y de la Información, 2020.

xviii, 182 p. - (Tecnologías de la información)

ISBN:

Investigación realizada gracias al programa

DGAPA - PAPIIT IT400118.

1. Inteligencia artificial - Procesamiento de datos. 2. Internet
de las cosas. 3. Archivos sonoros. 4. Big data. I. Rodríguez
Reséndiz, Perla Olivia, coordinadora. II. ser.

Diseño de portada: Oscar Fernando Arcos Casañas

Imágenes:

Envato Elements

(<https://elements.envato.com/es-419/>)

Primera edición, 2020

D.R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Ciudad Universitaria, 04510, México D.F.

Impreso y hecho en México

ISBN: En trámite

Publicación dictaminada

Aproximaciones a la inteligencia artificial y a los datos masivos en archivos digitales sonoros y audiovisuales

PERLA OLIVIA RODRÍGUEZ RESÉNDIZ
Universidad Nacional Autónoma de México

Desde la invención de la tecnología de la información, a principios del siglo XIX, nuestra capacidad de generación de datos nunca había sido tan intensa y constante. Cada año se produce más información digital que el anterior. El uso y manejo de grandes volúmenes de información digital es un signo contemporáneo.

Datos masivos, en inglés *big data*, es el término contemporáneo que alude a la identificación, procesamiento y recuperación de grandes volúmenes de información gracias al uso de la Inteligencia Artificial (IA) (Kusnetzky, 2014).

La extracción de información y conocimiento para desarrollar otras tareas es uno de los retos más importantes de la sociedad (Xindong Wu et al., 2014). De ahí que la aplicación de la IA en grandes volúmenes de información sea considerada una herramienta necesaria para la sistematización y análisis de datos.

La IA tiene una amplia gama de aplicaciones para enormes cantidades de datos. Por ejemplo, la presentación de proyectos como el algoritmo DeepMind de Google (DeepMind, 2019) que aprendió cómo ganar 49 juegos de Atari; así como la posibilidad de que los algoritmos puedan reconocer el lenguaje y patrones escritos a mano, describir la información de fotos y videos. Incluso, que el contenido de noticias se genere automáticamente. También se evidencian prometedoras aplicaciones en la salud al reducir los costos de la investigación clínica y ayudar a médicos y pacientes en la mejor toma de decisiones.

Además, en la biología y en la genética se observan soluciones de relevancia social. De hecho, la era ómica, es concebida como el periodo histórico en el que “se da una visión global de los procesos biológicos basada en el análisis de un gran volumen de datos... para estudiar organismos que son ahora desconocidos, así como sus funciones, todo a través de su rastro genético” (Monleon-Getino 2016, 433).

Además, la relevancia de los datos masivos y de la inteligencia artificial en la economía mundial es incuestionable. Para 2020, se ha estimado que el crecimiento de la industria de los datos a nivel mundial será de 739 mil millones de euros, que equivalen al 4% del producto interno bruto (PIB) y a 10 millones de empleos, directos e indirectos, que dependerán de este sector (Pellicer y Fariza, 2019). Los datos se han convertido en un activo clave para la economía (Monleon-Getino, 2016).

Por ello, la resistencia a la dependencia de empresas como Amazon, Google o Microsoft, en materia de almacenamiento y gestión de grandes volúmenes de datos ocupa la agenda política a nivel internacional. A finales del 2019, la canciller alemana Ángela Merkel declaró “muchas compañías han subcontratado [la gestión, el almacenamiento y el procesamiento de] sus datos a empresas estadounidenses... no digo que sea malo, sólo quiero decir que los productos de valor añadido que proceden de ahí con la ayuda de la inteligencia artificial crearán dependencias que no estoy segura de que sean buenas” (Pellicer y Fariza 2019, s/p). En el Parlamento Europeo se ha señalado que los datos y la inteligencia artificial son elementos para la innovación y para resolver problemas sociales, desde la salud, agricultura, hasta la seguridad. Por ello, se asignaron 20 millones de euros anuales para el desarrollo de la inteligencia artificial (Pellicer y Fariza, 2019).

La Unión Europea diseñó una estrategia sobre *big data* para apoyar y acelerar la transición hacia una economía basada en los datos en el espacio europeo. La economía basada en datos estimulará la investigación y la innovación en general, mientras que lleva a más oportunidades de negocio y a un aumento de la disponibilidad de los conocimientos y el capital, en particular para las pequeñas y medianas empresas (PYME) (Monleon-Getino 2016, 431).

Es sabido que la manipulación en el procesamiento de grandes volúmenes de datos de ciudadanos puede incidir en el manejo de la

toma de decisiones de la sociedad y poner a la democracia en riesgo (Helbing et al. 2018). También se ha evidenciado que en el diseño de herramientas de IA pueden incidir perspectivas subjetivas individuales o de grupo que excluyan a personas por raza, creencias, preferencias o bien afiliaciones ideológicas. Este es, sin lugar a dudas, un ámbito de investigación que debe ser abordado con mayor detalle para establecer los fundamentos éticos del uso de la IA en grandes volúmenes de contenidos digitales.

Estamos ante una revolución de los datos fenómeno que comprende el movimiento de los datos abiertos, el incremento en la colaboración pública (*crowdsourcing*), las nuevas tecnologías de la información, la inteligencia artificial y el internet de las cosas que están transformando a la sociedad (ONU, 2017) y no podemos tener una actitud pasiva, es necesario formular una perspectiva crítica del uso masivo de los datos a través de herramientas de IA.

ARCHIVOS DIGITALES

La incorporación de la tecnología digital en los archivos sonoros y audiovisuales modificó la noción de documento analógico a ítem, objeto y documento digital. Asimismo, transformó en ciclo de vida digital a los procesos documentales físicos e intelectuales, a través de los cuales se preservaron durante décadas los soportes analógicos. Además, confirió dinamismo y vitalidad a las instituciones de la memoria destinadas al resguardo de este tipo de colecciones. Por ello, la curaduría, como nuevo proceso intelectual se ha instaurado en los archivos para dar valor y visibilidad a los materiales que durante años sólo se conservaban. Con ello, la generación de nuevos contenidos y servicios de información que no habían sido imaginados ocupó el interés de los archivistas.

También, se dio origen a la construcción de archivos digitales, que pueden ser definidos como complejos sistemas de información que preservan cuantiosos volúmenes de contenidos digitales. Los archivos digitales, son un término polivalente que expresa la tecnología, un conjunto de procesos, técnicas y tecnologías empleadas en la preservación digital y además, puede ser considerada la institución de

la memoria que tiene como misión la salvaguarda de objetos digitales. Además, de forma coloquial se nombra como archivo digital a los registros digitales unitarios que se realizan en diversos formatos y lenguajes.

A pesar de las amplias posibilidades que ofrece el contexto digital para los archivos digitales, existen desafíos aún no resueltos. Algunos son derivados de la incorporación de nuevas técnicas y tecnologías digitales y otros, heredados del trabajo con soportes analógicos.

Uno de estos problemas es la continua acumulación de contenidos digitales (media y metadata) producto tanto de la digitalización como de la producción de documentos de origen digital.

Lo que significa que digitalizar no es preservar. La digitalización es el medio para transferir contenidos de un soporte analógico a uno digital. La preservación digital es una tarea continua que no se debe interrumpir; basada en una visión sustentable y de largo plazo. Por ello, una vez que se digitalizan los contenidos requieren ser migrados de forma sistemática, como parte de un plan de preservación digital sustentable.

A esta problemática se suma la constante y creciente producción de contenidos sonoros y audiovisuales de origen digital. Una gran parte de este tipo de materiales desaparece, por desconocimiento de los métodos adecuados de preservación y porque, en algunos casos, se producen en formatos de compresión y propietarios, condiciones que no son recomendadas para garantizar su permanencia a largo plazo. Por ello, reiteradamente se ha advertido que los materiales de origen digital afrontan un alto riesgo de pérdida, mayor incluso que el de los soportes analógicos, cuyos contenidos no han sido digitalizados.

De todos los tipos de materiales de información, las grabaciones sonoras y audiovisuales son las que requieren mayor espacio de almacenamiento digital. Aun cuando los soportes para el almacenamiento masivo digital evolucionan como resultado de la inercia del mercado y, cada vez se pueden conservar mayores volúmenes de contenidos digitales, su costo y rápida obsolescencia tecnológica limitan su adquisición. Sólo los archivos que cuentan con políticas de preservación digital sustentable y cuyos recursos económicos están garantizados a largo plazo pueden instalar robustos sistemas de almacenamiento masivo digital con visión de largo plazo.

El volumen de los documentos digitales sonoros y audiovisuales que deben ser preservados supera los límites imaginados por archivistas y documentalistas sonoros y audiovisuales. La producción de los materiales de origen digital es constante y exponencial. El tratamiento documental demanda de saberes, técnicas y tecnologías para preservar cuantiosas cantidades de contenidos digitales sonoros y audiovisuales. Más allá de las limitaciones económicas y tecnológicas, el crecimiento en el volumen de contenidos digitales trae consigo otro tipo de problemas asociados al ciclo de vida digital y a la permanencia del contenido digital.

INTELIGENCIA ARTIFICIAL Y DATOS MASIVOS (*BIG DATA*)

De 2003 a 2008, se evidenció el crecimiento de los centros de datos en el mundo. En 2011, se reportaron 509,147 centros de datos en el mundo (Miller, R. 2011). En esta proyección se incluyen los centros de datos que preservan contenidos sonoros y audiovisuales. La incorporación del big data y de la inteligencia artificial en los archivos sonoros y audiovisuales es reciente si se considera que las primeras aplicaciones de la IA comenzaron en los años cincuenta del siglo pasado y el crecimiento del volumen de contenidos digitales es un rasgo, como se ha señalado, que caracteriza a la era contemporánea.

Los archivos sonoros y audiovisuales que han digitalizado sus colecciones o están llevando a cabo este proceso, acogen estas herramientas como un medio para resolver problemas asociados a la identificación y a la recuperación de la información. Debido al tiempo y complejidad que demanda el proceso de catalogación sonora y audiovisual, en muchos casos, sólo se cuenta con los datos que se pueden incorporar en la creación de inventarios. La catalogación es una tarea compleja porque difícilmente un archivista o documentalista puede asentar con precisión la amplia gama de materias que se resguardan en los archivos. Además, el grado de error en el asentamiento de los datos persiste a través de faltas de redacción y de ortografía, entre otras. Frente a este panorama, el uso de herramientas para identificar, transcribir y recuperar contenidos de forma automatizada, significa una alternativa en las tareas de preservación digital que deben afrontar los archivos.

En los primeros años del siglo XXI, se han puesto en marcha las primeras aproximaciones teóricas y desarrollos tecnológicos de la IA en grandes volúmenes de datos digitales que se resguardan en archivos sonoros y audiovisuales. Los resultados de las experiencias pioneras son muy valiosos. Como sucedió con las primeras experiencias de digitalización de archivos, el aprendizaje basado en el ensayo-error ha sido una pieza fundamental en el desarrollo y asentamiento de la digitalización como proceso tecnológico para transferir contenidos grabados en soportes analógicos a digitales.

Los científicos de las ciencias de la información, bibliotecología y archivología tienen, en el estudio e investigación de la inteligencia artificial y el big data, un ámbito emergente de conocimiento. Es necesario determinar los efectos sociales y culturales del big data y la inteligencia artificial. Además, es necesario conocer el impacto que estas tecnologías tendrán en las instituciones de la memoria que cada día incrementan el volumen de ítems digitales en bibliotecas, archivos y museos.

En este sentido, el libro *Inteligencia artificial y datos masivos* en archivos digitales sonoros y audiovisuales ofrece discusiones teóricas sobre el tema y presenta estudios de caso, que dan cuenta de los derroteros de los archivos sonoros y audiovisuales que han sido pioneros en la utilización de la IA en grandes volúmenes de datos digitales.

El capítulo “El papel de la inteligencia artificial en la Industria 4.0”, escrito por el Dr. Humberto Sossa, presenta el marco conceptual y el estado del arte de este ámbito del conocimiento que, de acuerdo con el autor, será detonante de la quinta revolución industrial, de la economía cíclica. El trabajo sitúa el desarrollo histórico de la IA a nivel internacional y ubica la posición de México, lo que constituye una aportación significativa si se considera que alentar la investigación científica en este campo incide en el desarrollo tecnológico, en la innovación y en la generación de nuevos mercados.

A continuación, la doctora Lily Díaz presenta, desde la perspectiva antropológica, el capítulo “Ontología y narrativa en los sistemas de inteligencia artificial, una mirada antropológica”. Ofrece una reflexión teórica en relación con los grandes volúmenes de datos y su recuperación e interpretación a través de la inteligencia artificial. Esta visión establece una primera discusión en torno a la relación del big

data y la inteligencia artificial que señala la necesidad de contar con métodos que permitan trabajar de forma creativa con datos de forma responsable e inclusiva.

El uso de la IA en grandes volúmenes de datos es, sin lugar a dudas, una fuente de progreso e innovación; no obstante, al mismo tiempo establece aspectos jurídicos que es necesario considerar. En el apartado “Big Data Between Privacy and Copyright” de Amelia Bongarzone, se exploran, desde una perspectiva jurídica, las posibilidades que ofrece el big data para elaborar modelos interpretativos, analíticos e incluso predictivos de acontecimientos y comportamientos humanos, difíciles de imaginar hasta hace poco, y que llaman la atención del derecho privado y suscitan la reflexión sobre las implicaciones que pueden afectar a la sociedad. La visión jurídica va más allá del ámbito patrimonial.

Antonella Bongarzone en “Una estrategia para gestionar los big data en archivos sonoros y audiovisuales”, introduce el tratamiento de documentos dinámicos, versus el estático. Establece que la catalogación, con base en las normas internacionales establecidas, es portadora de información sobre el contenido sonoro o visual. Más allá de este propósito también se reflejan el contexto histórico, sociológico, antropológico y cultural en que se produce dicho registro. Bajo esta consideración, el uso potencial de del big data, desde la perspectiva de las ciencias sociales, ofrece la posibilidad de ampliar el análisis y profundizar las descripciones superficiales con datos profundos del contexto del documento que se derivan precisamente del contexto histórico. La autora se centra en el significado y el valor del dato para que los grandes volúmenes de información no se conviertan en un mar incontrolado, sino que sean verdaderos vehículos de conocimiento.

En el capítulo “Using Computational Tools and Experts to Improve Access to Digital Media Archives”, escrito por Karen Cariani y David O Ives, se exponen las posibilidades de uso del aprendizaje automático, en inglés Machine Learning (ML) y la inteligencia artificial, en el American Archive of Public Broadcasting (AAPB), de Estados Unidos. El trabajo expone a partir de un estudio de caso la necesaria colaboración que debe existir entre las instituciones que preservan grandes cantidades de datos digitales sonoros y audiovisuales y los científicos que desarrollan y necesitan probar herramientas para el manejo de

grandes volúmenes de datos. Se sostiene la necesaria colaboración para mejorar la IA y el ML para obtener más y mejor calidad en los datos que se resguardan en archivos y bibliotecas.

Virginia Bazán-Gil ofrece la experiencia de la Radio y Televisión Española (RTVE) en la incorporación de la inteligencia artificial como una herramienta para dar visibilidad a los archivos. El capítulo “Reflexiones en torno a la inteligencia artificial: el caso de los archivos de RTVE”, presenta la experiencia de la incorporación de la IA para la generación automática de metadatos a través de la transcripción automática de voz a texto, así como de la segmentación y agrupación por hablantes y caras para el uso de contenidos en la producción de RTVE. El texto es una lección, necesaria para los archivos de radio y televisión, que, una vez concluido el proceso de digitalización de sus colecciones, afrontan el desafío de ensanchar las posibilidades de acceso. El apartado es muy claro y aleccionador, describe los alcances y limitaciones de la IA en un medio público de radio y televisión. Expresa cómo la incorporación de este tipo de tecnologías no debe ser una tarea aislada, sino resultado de la sinergia entre las diferentes áreas del medio.

En el apartado “Catalogación automática: más cerca que lejos”, escrito por Alejandro Álvaro Ramírez Acosta y Juan Carlos Murillo Pabón, se presentan los avances del Laboratorio de Radio e Inteligencia Artificial de Radio Televisión Nacional de Colombia (RTVC), que trabaja con los documentos del archivo sonoro de Señal Memoria. La aplicación de la IA en la experiencia colombiana está destinada a apoyar el proceso de catalogación de los archivos.

La evolución tecnológica es el punto de partida del capítulo “Apuntes sobre la Inteligencia Artificial y su aplicación en los archivos sonoros” de Georgina Sanabria. En este trabajo se determina, desde una perspectiva histórica, la evolución de la IA y sus posibles aplicaciones en los archivos sonoros. La autora destaca que, aun cuando el término IA es utilizado de forma recurrente en entornos comerciales, no se pueden omitir las posibilidades que ofrece esta tecnología en la búsqueda de soluciones a problemas sociales, uno de los cuales compete a la preservación de la memoria sonora de la humanidad.

La complejidad que implica la descripción de grandes volúmenes

de información afecta también a los archivos de fotografías. Esta problemática recurrente a los archivos sonoros y audiovisuales, impacta también en archivos de imagen fija como son los fotográficos. De ahí que en el capítulo “Identifying Metadata Elements in Photographic Repositories by Semantic Segmentation of Images with Deep Learning”, escrito por Sinaí López-Castillo, Isnardo Reducindo y Francisco Benita, se presente un método para la descripción automatizada de fotografías a través del uso de algoritmos de Deep Learning (DL).

El último capítulo “Propuesta metodológica y de análisis computacional para identificar el proceso fotográfico en fotografías históricas del siglo XIX y XX”, de la autoría de Gustavo Lozano y Rodrigo Colín Rivera, se ofrece una metodología para identificar a partir del uso de la IA la técnica fotográfica. Con ello, se busca proponer una herramienta a disposición de archivistas para la catalogación de fotografías.

Este libro es un trabajo pionero en el uso de la inteligencia artificial y los datos masivos en archivos digitales sonoros y audiovisuales, su lectura puede contribuir a comprender los alcances actuales y futuros de estas tecnologías en la preservación digital sustentable.

REFERENCIAS

DeepMind. (2019). <https://deepmind.com/>

Kusnetzky, D. (2014). *What is «Big Data?»* ZDNet. <https://www.zdnet.com/article/what-is-big-data/>.

Monleon-Getino, A. (2016). *El impacto del Big-data en la Sociedad de la Información. Significado y utilidad*. *Historia y Comunicación Social*, 20(2), 137-155. https://doi.org/10.5209/rev_HICS.2015.v20.n2.51392.

Pellicer y Fariza. (2019). *Europa ultima un plan para dar la batalla en el negocio de los datos | Economía | El País*. https://elpais.com/economia/2019/11/16/actualidad/1573926886_318836.html.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, & Wei Ding. (2014). Data Mining with Big Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 26, núm. 1, JANUARY 2014.

Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión especializada, Angélica Valenzuela y Valeria Guzmán González; revisión de pruebas, Valeria Guzmán González; formación editorial, Oscar Fernando Arcos Casañas. 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 2020.