El Futuro del Pasado: Algunas reflexiones sobre el desarrollo de Inteligencia Artificial en la Historia Novohispana, la Arqueología Histórica, y la descolonización tecnológica.

Dra. Patricia Murrieta-Flores. Catedrática en Humanidades Digitales, Departamento de Historia, Universidad de Lancaster, Reino Unido.

Introducción

Caminando por los pasillos de mi universidad y de camino a uno de los seminarios de investigación en Humanidades Digitales (HD) que celebramos en Lancaster, puedo percibir el entusiasmo de estudiantes e investigadores por conocer las últimas novedades y aplicaciones de técnicas computacionales, en esta ocasión, a manuscritos medievales y evidencias arqueológicas en torno a la historia de la Sicilia Medieval (Fernández-Aceves, 2017). Las Humanidades Digitales están presentes desde hace varias décadas, pero todavía da la sensación de ser un campo nuevo. La razón de esto podría ser porque a medida que los enfoques computacionales y la Inteligencia Artificial evolucionan, y herramientas como ChatGPT-4 y DALL-E empiezan a estar al alcance de todo el mundo, percibimos que si bien antes las tecnologías digitales eran omnipresentes, ahora más que nunca están configurando activa y dinámicamente nuestro mundo, y esto incluye disciplinas académicas enteras.

Cuando me preguntan ¿qué son las Humanidades Digitales?, suelo dar una respuesta sencilla y amplia: Las HD ponen al centro la aplicación de teorías y métodos de las Ciencias de Datos y de la Computación para resolver preguntas de Humanidades. La razón por la que doy una definición tan amplia es porque la complejidad del campo así lo requiere. La pregunta exige cuestionarse ¿qué son las Humanidades? ¿Y cuáles son los puntos en común entre estas disciplinas que nos permitan pensar en teorías, métodos y enfoques digitales aplicables a todas ellas? Una mirada al campo desde revistas establecidas como *Digital Humanities Quarterly* revela el enorme panorama de las HD, y por lo tanto, estudiantes y colegas a menudo me preguntan cuál es un buen punto de entrada a la disciplina. A lo largo de mi carrera académica, he tenido la suerte de trabajar con expertos en Historia, Literatura y Arqueología colaborando en la creación de métodos computacionales que permitan responder a algunas de las intrincadas preguntas que éstas disciplinas se plantean. Mis intereses se han alineado principalmente con la geografía y la creación de métodos, y aunque no son las únicas líneas de investigación en HD, este interés surgió en mi caso debido a mi carrera como arqueóloga.

A principios de los años 2000 al final de la licenciatura, hice mi primera incursión en investigación a través de la historia y la arqueología colonial de la industria azucarera. Como estudiante, sabía muy poco de las formas en que las innovaciones de la informática podían combinarse con la historia y la arqueología, y que la investigación interdisciplinaria entre estos campos ya sucedía. Mientras que las Humanidades Digitales se iniciaron en la década de 1940 en el ámbito anglo-parlante, a finales del siglo XX en México, la práctica consistente de la arqueología histórica a pesar de sus orígenes en la década de 1960, todavía se sentía poco establecida[1]. Esto está relacionado probablemente, al fuerte énfasis en la arqueología prehispánica por parte de instituciones clave como el INAH y la ENAH tanto en la investigación como la docencia. Además, el país estaba ciertamente aún lejos de recurrir a enfoques computacionales de manera regular en cualquiera de los campos de las humanidades.

[1] El primer Congreso Nacional de Arqueología Histórica en México tuvo lugar en 1996 en Oaxaca. Véase la reflexión de Elsa Hernández Pons (2000) sobre el estado de la cuestión en aquel momento.

También eran relativamente escasos los proyectos altamente interdisciplinarios. Sin embargo, la idea y el apetito por estas cosas estaban ya presentes. Las investigaciones emblemáticas en arqueología histórica de académicos como Elsa Hernández Pons, Jaime Litvak King[2], Elsa Malvido, Pilar Luna Erreguerena, Patricia Fournier, Mónica Lugo Ramírez, José Alberto Aguirre Anaya, y Flor Trejo Rivera, entre muchos otros, abrirían el camino para generaciones posteriores como la mía, y también se estaban gestando proyectos digitales pioneros enfocados a las tecnologías de la información y la divulgación. Por ejemplo, la presentación digital de Marc Thouvenot del Códice Xolotl, serviría de base para Tlachia (https://tlachia.iib.unam.mx/). Llamada así por el verbo náhuatl "ver o mirar", Tlachia constituye una base de datos única de elementos pictóricos procedentes de códices coloniales nahuas. El método de análisis aplicado en este conjunto de datos se basa principalmente en el sistema de investigación y clasificación ideado por Joaquín Galarza y su grupo de investigación. A pesar de tener innumerables objeciones infundadas en el ámbito anglo-parlante, sigue siendo probablemente el método más utilizado de clasificación de documentos pictóricos nahuas, y en términos de sistemas de información, se está convirtiendo en la base para crear nuevos motores de búsqueda de fuentes históricas pictográficas. Aunque Tlachia funciona como sitio Web y desde el punto de vista de la interfaz de usuario es ciertamente un producto de su tiempo, la información que proporciona no tiene paralelo. También surgido o inspirado en este trabajo, el Gran Diccionario Náhuatl (https://gdn.iib.unam.mx/), sería otra iniciativa sin parangón y hoy en día es ampliamente utilizada por estudiosos y miembros del público en general. Este recurso recopila miles de términos y palabras en náhuatl en su contexto a partir de variantes y sus traducciones al español que abarcan desde 1547 hasta 2002. Otro proyecto digital de importancia es Amoxcalli (https://www.amoxcalli.org.mx/). Financiado por el CONACyT y bajo la dirección de Luz María Mohar Betancourt, Amoxcalli lograría lo que la Biblioteca Nacional de Francia en su momento no había conseguido: publicar en línea una parte sustancial de los acervos del Fondo Mexicano de esa biblioteca, acompañados de sus transcripciones. Como en el caso anterior, la actual interfaz en línea podría parecer básica a los ojos acostumbrados a las nuevas tecnologías, sin embargo, en términos de historia de las HD, éste sería sin duda un ejemplo a seguir para los posteriores practicantes en México. La información producida por estos proyectos sigue siendo utilizada por innumerables estudiosos, incluyendo mis grupos de investigación. Existen muchos otros proyectos de finales del siglo XX, como la Biblioteca Digital del Pensamiento Novohispano (http://www.bdpn.unam.mx/), el Diccionario del Español de México (https://dem.colmex.mx/), que también podrían pensarse como los pioneros de la historia digital Novohispana y que por falta de espacio no puedo mencionar ampliamente. Sin embargo, no cabe duda de que todos ellos serían los precursores de proyectos actuales como el Corpus Electrónico del Español Colonial Mexicano (https://www.iifilologicas.unam.mx/coreecom/), el Lienzo Digital de Tlaxcala (https://lienzodetlaxcala.unam.mx/), los Códices de México del INAH (https://www.codices.inah.gob.mx/pc/index.php), la edición digital del Códice Mendoza (https://codicemendoza.inah.gob.mx/), el Diccionario Náhuatl de Wired-Humanities (https://nahuatl.wired-humanities.org/), y proyectos de participación pública muy exitosos como Noticonquista (https://www.noticonquista.unam.mx/).

Si bien estos son desarrollos notorios, en términos de epistemología, creo que es importante entender las complejas capas de las HD como campo. Los ejemplos presentados arriba pueden

[2] Gracias a sus innovadores trabajos sobre modelado estadístico, Jaime Litvak está considerado uno de los pioneros mundiales de la arqueología computacional.

considerarse un tipo de las muchas manifestaciones que este tiene en América Latina, y que muchas veces toma nombres como 'historia digital', 'arqueología computacional', 'lingüística histórica digital', etc. La gran variedad de proyectos y disciplinas dentro de las HD, puede verse también en la conformación de la Red HD (http://humanidadesdigitales.net/), formada por académicos de varias instituciones y campos como la Filosofía, Historia, Lingüística y Bibliotecología, entre muchas otras. Testimonio también de la diversidad de las HD latinoamericanas fue la Conferencia Internacional de Humanidades Digitales de 2018, celebrada en la Ciudad de México. Esta fue la primera conferencia de ADHO (https://adho.org/) organizada fuera del Norte Global. En cuanto a las prácticas dentro del campo, la creación de bases de datos y la difusión de información histórica o arqueológica es necesaria e importante, pero como se ha dicho, las HD a menudo toman otras formas. Estas pueden incluir el desarrollo de visiones teóricas críticas hacia los enfoques computacionales actuales, la creación de metodologías, el diseño de tecnologías o infraestructura, o la creación de software para el estudio de evidencia histórica y arqueológica, entre muchas otras. Los proyectos que presentaré en la penúltima sección ejemplifican algunas de estas formas, y éstos, están también relacionados con aquellas primeras investigaciones que realicé al principio de mi carrera académica sobre la industria azucarera en época colonial.

Hacia 2004 y como parte de un proyecto del INAH dirigido por Elsa Hernández Pons, excavamos por primera vez una hacienda azucarera en el estado de Guerrero. La Hacienda de Tecoyutla tiene una historia fascinante, y como parte de mi investigación visité el Archivo General de la Nación (AGN) en innumerables ocasiones, buscando información sobre el sitio, su historia, y todo aquello que pudiera darme una idea de su cautivante desarrollo. Mientras estaba en el archivo, surgieron muchas preguntas que no pude resolver con los sistemas existentes. Leyendo un documento en el que se mencionaba un topónimo desconocido y que no podía localizar, me pregunté dónde podría haber estado y cómo encontrarlo. ¿Cuántas y qué otras haciendas se construyeron alrededor de los mismos años que Tecoyutla? ¿Existía una red socioeconómica entre haciendas o trapiches en esta región? ¿Qué se podría encontrar sobre los procesos de producción en los otros miles de documentos que estaban aparentemente a mi disposición, pero en realidad, muy lejos de mi alcance como una investigadora individual? Esta sensación de imposibilidad se debía no sólo al gran volumen de documentos del archivo, sino también al hecho de que, en aquel momento, si quería responder a preguntas tan sencillas, tendría que perseguir y localizar documentos no sólo en el AGN, sino posiblemente en otros archivos del extranjero como el Archivo General de Indias, viajar (y ahorrar dinero) para consultarlos en persona, transcribir cientos, si no miles de páginas, e identificar la información de mi interés en todos ellos. Confrontada con los sistemas de búsqueda del AGN, veía frente a mí, el trabajo de toda una vida. Creo que fue la primera vez que soñé con las posibilidades de un motor de búsqueda basado en cartografía o mapas que me ayudara mágicamente en mi búsqueda. En los años siguientes, exploraría el uso de los Sistemas de Información Geográfica y me convencería de que este tipo de tecnologías eran el futuro de la arqueología histórica. Apoyada por una de las becas que ofrece el CONACyT, esto me llevó a hacer una maestría y un doctorado en un campo llamado en el Reino Unido "archaeological computing". El nombre hace referencia a sus raíces en la combinación de métodos informáticos con la arqueología, y esta especialización me llevaría a trabajar en un proyecto europeo llamado 'The Spatial Humanities'. La premisa de este proyecto era explorar la improbable idea de utilizar Sistemas de Información Geográfica (SIG) con información textual. Con sus orígenes en las ciencias medioambientales y naturales, los SIG se utilizaban exclusivamente con datos producidos en estos ámbitos, normalmente derivados de

trabajo de campo o contextos similares, y nunca se había intentado combinar esta herramienta con fuentes textuales. El proyecto de Spatial Humanities llevó a cabo dos estudios de caso con el objetivo de incorporar corpus textuales con SIG. La idea básica era identificar las geografías descritas en uno o varios textos, cartografiarlas en SIG, y desarrollar la capacidad de explorar lo que se decía sobre esos lugares mediante una combinación de análisis espacial y crítica literaria (Gregory y Copper, 2011). En uno de los casos se utilizó un corpus de textos literarios que abarcaron del siglo XVII al XIX del emblemático Distrito de los Lagos inglés, para explorar cómo percibían y representaban sus geografías los poetas y escritores del pasado (Donaldson et al., 2015), y en otro, se exploraron las enfermedades y la mortalidad en Inglaterra y Gales durante época Victoriana, utilizando un amplio corpus de informes históricos gubernamentales. Esta investigación llevaría a la creación de un método llamado Análisis Geográfico de Textos (Murrieta-Flores et al., 2015), y al posterior diseño de dos proyectos de HD centrados en la creación de métodos computacionales, conjuntos de datos, software, y enfoques para el estudio de la historia colonial de Nueva España.

El desarrollo de la inteligencia artificial para la Historia Novohispana y la investigación en Arqueología Histórica

A medida que mi trabajo fue avanzando en las Geohumanidades, empecé a interesarme por la Inteligencia Artificial (IA). La IA también es un campo muy amplio, y debido a algunas connotaciones problemáticas más al margen de la ciencia ficción que de la ciencia real (incluidas las ideas en torno al transhumanismo), creo que el término 'aprendizaje automático por computadora -*Machine Learning*' describe con mayor precisión lo que el campo hace. El aprendizaje automático por computadora es una subdisciplina de la inteligencia artificial que implica el desarrollo de algoritmos y modelos que permiten a las computadoras aprender de los datos, sin estar explícitamente programadas para realizar una tarea específica. Los algoritmos de aprendizaje automático están diseñados para identificar patrones en los datos, hacer predicciones o tomar decisiones basadas en esos datos y mejorar continuamente su rendimiento a lo largo del tiempo ajustando sus parámetros o aprendiendo de nuevos datos (Zhou, 2021). El objetivo del aprendizaje automático es permitir que las computadoras aprendan y mejoren su rendimiento en una tarea específica, sin intervención humana, aprovechando técnicas estadísticas y computacionales. Esto tiene numerosas aplicaciones prácticas en campos como el procesamiento del lenguaje natural, la visión por computadora, el reconocimiento automático del habla, y el análisis predictivo (Jordan y Mitchel, 2015). Aunque no descarto el potencial de aplicaciones en el uso, por ejemplo, del reconocimiento automático del habla en etnografía y otros campos antropológicos, en el caso de la arqueología histórica, los dos campos de IA que resultan de extrema relevancia son el Procesamiento del Lenguaje Natural y Visión por Computadora.

El Procesamiento del Lenguaje Natural -*Natural Language Processing* (PLN) es un campo de estudio que se centra en la interacción entre el lenguaje humano y las computadoras. Consiste en desarrollar algoritmos y modelos que les permitan analizar, comprender y generar lenguaje natural (Chowdhary, 2020). Ejemplos de sus aplicaciones son los asistentes automatizados como Siri de Apple, Amazon Alexa, Google Assistant o Cortana de Microsoft, y más recientemente ChatGPT-4, entre muchas otras. El PLN puede utilizarse en la investigación histórica para analizar grandes volúmenes de textos históricos como documentos, cartas, y artículos de periódicos con el fin de extraer información y patrones significativos. Las técnicas de PLN también pueden ayudar a los historiadores a identificar o extraer automáticamente "entidades clave", incluyendo la mención de

personas, lugares, eventos y otros conceptos complejos en miles de páginas o documentos en cuestión de segundos (Sporleder, 2010; Won et al., 2018; Hutchinson, 2020). A partir de ahí, pueden surgir muchos tipos de análisis y plantearse preguntas complejas. Por ejemplo, el investigador puede querer investigar la red social y las relaciones entre personajes históricos, o identificar la mención de determinados recursos o acontecimientos en un documento, en una gran colección de libros o fuentes históricas. Otros métodos, como el análisis de sentimientos (*sentiment analysis*), pueden dar una idea del tono emocional de textos históricos como cartas y diarios, ayudándonos a comprender mejor las actitudes y emociones de la gente en el pasado. El PLN también puede utilizarse para identificar patrones lingüísticos en fuentes, como marcadores lingüísticos de clase social o variantes regionales o dialectales. Estas son sólo algunas de las aplicaciones de las herramientas en este campo, que cada vez ayuda más a los investigadores a analizar grandes volúmenes de textos históricos de maneras novedosas.

Otro campo que me parece cada vez más importante para la historia y la arqueología es la visión por computadora -*Computer Vision*. Este es un campo de la inteligencia artificial y la informática que se centra en capacitar a las computadoras para interpretar, comprender y analizar imágenes digitales y vídeos del mundo real (Shapiro y Stockman, 2001). Implica el desarrollo de algoritmos, modelos y sistemas capaces de extraer información relevante de los datos visuales, como el reconocimiento de objetos, la reconstrucción de escenas, el análisis del movimiento y el procesamiento de imágenes. El objetivo de la disciplina es permitir que las máquinas repliquen las capacidades de la visión humana y que puedan realizar tareas que requieren percepción visual, como la navegación autónoma, el reconocimiento facial, el diagnóstico médico, y la vigilancia policial (Voulodimos et al., 2018). Aunque muchos de estos usos son controvertidos, y diré más al respecto en la sección final sobre los problemas que existen actualmente con este tipo de tecnología en términos de colonialismo y sesgos en los conjuntos de datos que se han utilizado tradicionalmente, en el caso de la historia novohispana esta tecnología nos está permitiendo abrir el acceso a los archivos de una manera sin precedentes. Por ejemplo, estamos tomando grandes colecciones de documentos históricos y llevando a cabo la transcripción automatizada de los mismos. Los algoritmos utilizados también pueden entrenarse con imágenes históricas como mapas y pinturas, para reconocer objetos específicos representados en ellas, como edificios, personas, rasgos del paisaje, animales, símbolos, iconografía, colores, características especiales y glosas, entre muchos otros. Además, a través de estos sistemas se puede clasificar estos elementos según el interés del investigador. Estas técnicas también pueden tener otros muchos usos en el estudio de la cultura material y la historia moderna. Diego Jiménez Badillo, por ejemplo, ofrece en este volumen un interesante estudio de caso sobre la clasificación de objetos arqueológicos con este tipo de método, y estas técnicas también pueden utilizarse para reconstruir escenas históricas a partir de imágenes o vídeos, permitiendo a los investigadores visualizar el contexto en el que tuvieron lugar los acontecimientos históricos. Por ejemplo, utilizando fotos del siglo XIX y principios del XX, un investigador podría adentrarse en el estudio de paisajes históricos y su evolución.

Estos métodos también pueden combinarse con técnicas de otros campos y diferentes enfoques computacionales, como los Sistemas de Información Geográfica, la Lingüística de Corpus, y los Datos Abiertos Vinculados. Puesto que no hay lugar sin historia ni historia sin lugar, las teorías del espacio, la geografía y el paisaje han estado habitualmente en la vanguardia de los estudios históricos y arqueológicos. Como sistema informático que permite a los usuarios recopilar, almacenar, manipular, analizar y visualizar datos referenciados geográficamente, era natural que los SIG se

convirtieran en un dispositivo habitual en el conjunto de herramientas del arqueólogo. Sin embargo, para dar ese "salto" a otros campos de las humanidades como la historia y la literatura, los SIG tenían que trabajar también con el tipo de evidencia más habitual en estas disciplinas, es decir, fuentes escritas. Esto solo fue posible a través de la integración de los SIG con métodos de la Lingüística de Corpus (Murrieta y Gregory, 2015). La investigación en este campo, implica el análisis de datos lingüísticos típicamente de grandes colecciones electrónicas de texto, conocidas como corpus, e incluye análisis computacionales como concordancia, palabras clave, y análisis de colocación (McEnery, 2019). Estos métodos permiten a los investigadores identificar patrones y tendencias en el uso del lenguaje, como la frecuencia de palabras o frases específicas, la distribución de estructuras gramaticales y el uso del lenguaje en diferentes contextos. El resultado de la combinación entre SIG, Lingüística de Corpus y PLN, permitió crear el ya mencionado Análisis Geográfico de Textos (GTA por sus siglas en inglés) cómo método y software (Murrieta-Flores et al., 2022). Esto fue posible en el contexto del proyecto *Digging into Early Colonial Mexico* y utilizando el corpus de las Relaciones Geográficas de Nueva España de Siglo XVI, como se explica más adelante. Con este método y herramienta el investigador puede identificar automáticamente en miles de páginas de documentos históricos información como: los topónimos que menciona la fuente; la mención de una entidad o concepto particular de su interés (como nombres propios, fechas, eventos, actividades, animales, etc.) y su contexto; si es que esta entidad o concepto está asociado a una geografía específica; y qué se dice de ellos. Por ejemplo, una investigadora, podría querer identificar todas las plantas descritas en la colección completa de las Relaciones Geográficas, lo que se dice de ellas, y los lugares de los que proceden, incluidas sus coordenadas para crear un mapa o realizar análisis espaciales posteriores. Aunque esto es ciertamente posible de forma manual, hay que considerar que las *Relaciones* se compilan en 12 volúmenes contando con cientos de páginas y casi 3.000.000 de palabras. Con el GTA software, podemos realizar esta tarea en cuestión de segundos y con sólo unos clics (FIGS. 1 y 2).

Fig. 1. El Geographical Text Analysis software se compone de 5 herramientas: a) el Corpus Tool muestra todos los documentos en los que se está realizando una consulta y el texto anotado; b) la sección de Filters permite la búsqueda de términos específicos de interés, utilizando palabras clave, seleccionando una combinación de entidades y etiquetas, o búsquedas complejas con expresiones booleanas, etc; c) el Gazetteer tool muestra los topónimos que la fuente histórica menciona; d) el Context Tool muestra los resultados del termino, palabra clave, entidad o frase buscada, con su contexto a la izquierda y a la derecha en la oración. Esto incluye resultados en la búsqueda de todo el corpus; e) el Annotation Information Tool muestra los detalles aceca de la anotación/palabra/frase de interés, incluyendo su entidad o etiqueta, contexto, e información geográfica.[3]

Fig. 2. En este ejemplo, se buscó la mención de plantas en todo el corpus de las Relaciones Geográficas y lo que se dice de ellas en su contexto. Se encontraron 6240 menciones en total. Sin embargo, tal vez estemos interesados en las geografías del “henequén”. El ejemplo muestra las 130 menciones de “henequén” encontradas en su contexto y como ésta mención en específico está conectada con el sitio de Cuautla en las Relaciones de Antequera.

[3] Todas las ventanas están conectadas y son interactivas, de forma tal que las búsquedas no están limitadas al Filter Tool. Por ejemplo, si uno quiere buscar lo que se dice de un lugar en particular, se puede utilizar el mapa dibujando un polígono alrededor de las geografías que le interesan. Esto da como resultado todas aquellas oraciones que mencionan dichos topónimos; o uno puede resaltar en la ventana de lectura una cadena de palabras, por ejemplo una expresión lingüística particular, y obtener todas buscando en todo el corpus aquellas oraciones que contienen dicha expresión.

También podemos exportar esta información a un formato de hoja de cálculo que luego podemos llevar a un programa de cartografía como los SIG, Google Maps, o Earth, o a cualquier otro tipo de herramienta digital (FIG. 3). Con este método también podemos hacer preguntas complejas. Por ejemplo, al solicitarle que recupere toda la información relacionada con la salud y la enfermedad en los 12 volúmenes de las Relaciones Geográficas, actualmente estamos utilizando este software para comprender mejor cómo se dieron las enfermedades en toda la Nueva España y cómo estaban cambiando y registrándose las prácticas médicas a finales del siglo XVI. Este tipo de análisis, sin embargo, no está limitado a una colección de fuentes o a un tipo individual de consulta. Imaginemos las posibilidades. En un futuro no muy lejano, podremos consultar por ejemplo, todas las crónicas de la conquista, o todas las fuentes producidas a lo largo de siglos completos a la vez. Esto abre la oportunidad de identificar patrones y analizar cambios sociales, culturales, políticos y económicos a lo largo del tiempo, a una escala que antes era imposible.

Fig. 3. La herramienta GTA puede exportar la consulta a formatos como .xls o .json que permiten otras formas de análisis y exploración desde el punto de vista geográfico, cuantitativo, o lingüístico, entre muchos otros. En este ejemplo se muestra la extracción de miles de menciones y cientos de términos relacionados con la salud, la enfermedad y los remedios registrados en las Relaciones Geográficas de Siglo XVI.

En el caso de los Datos Abiertos Vinculados -*Linked Open Data*, su objetivo es permitir el intercambio y la integración de datos entre diferentes dominios y aplicaciones, fomentando la colaboración y la innovación en una amplia gama de campos. Esto incluye el uso de vocabularios y ontologías estandarizados para describir objetos en conjuntos de datos y proporcionar enlaces a otros datos relevantes (Yu, 2011; Lausch et al., 2015). Esto permite crear una red de datos descentralizada e interconectada, en la que diferentes conjuntos de datos pueden combinarse y consultarse fácilmente, y en la que máquinas y personas pueden realizar razonamientos y análisis inteligentes. Esto puede ser de gran utilidad para el análisis y exploración de datos en el patrimonio histórico y cultural (Pereda, 2019). Pensemos en el ejemplo anterior sobre Visión por Computadora y la identificación automatizada de características en imágenes históricas. Siguiendo los pasos de Tlachia, estamos realizando en este momento experimentos de clasificación de imágenes con los mapas de las Relaciones Geográficas y otras colecciones, buscando conectar esta información con Datos Abiertos Vinculados (FIG. 4). Esto tiene el potencial de resultar en grandes bases de datos y en el desarrollo de motores de búsqueda que sean capaces de conectar la información de las imágenes históricas y los textos, pero también con datos arqueológicos y entre conjuntos de datos en fuentes dispares y colecciones dispersas por todo el mundo. Este tipo de desarrollo facilitará el análisis y la comparación de miles de fuentes que ayudarán a comprender fenómenos como el desarrollo de la escritura y las tradiciones cartográficas de Mesoamérica, así como sus cambios a lo largo del tiempo con la introducción de conceptos espaciales europeos.

Fig. 4. Ejemplo de nuestros primeros experimentos en identificación automática de elementos en mapas coloniales.

Digging into Early Colonial Mexico y Unlocking the Colonial Archive

Como se puede ver, la inteligencia artificial y los enfoques computacionales ofrecen caminos innovadores para la arqueología histórica, y nuestro objetivo en los últimos 10 años, ha sido aprovechar estas tecnologías para crear no sólo nuevos conjuntos de datos que tengan un impacto sustancial en investigación, sino también para desarrollar nuevos enfoques, herramientas y métodos con tecnologías que pueden ayudar al avance de la historia colonial de México y América Latina. Esta

fue la misión de "Digging into Early Colonial Mexico: A large-scale computational analysis of 16th century historical documentsI" (https://www.lancaster.ac.uk/digging-ecm/) (DECM). Este proyecto contó con el apoyo de la Plataforma Transatlántica de Humanidades y Ciencias Sociales, y fue financiado por ESRC-Reino Unido, CONACyT-México y FTP-Portugal. Hemos descrito ampliamente el proyecto en otro lugar (Murrieta-Flores, Jiménez Badillo y Martins, 2022) por lo que bastará con decir aquí que, centrándonos en las Relaciones Geográficas como una colección histórica de mayor escala, nuestros objetivos fueron seguir desarrollando el Análisis Geográfico de Textos, crear un modelo de PLN para la anotación de estos documentos históricos, y crear el primer nomenclátor o diccionario geográfico histórico digital del siglo XVI de la Nueva España. Entre los principales resultados de este proyecto, además del software GTA ya cubierto, se creó una colección completamente anotada de las Relaciones Geográficas que ahora sirve para la minería de textos y otros análisis. Las anotaciones se basan en 18 entidades diferentes que identifican en todo el corpus palabras y conceptos relacionados con las categorías de persona, fecha, institución, lugar, animal, actividad, planta, alimento, recurso natural, artefacto cultural, arquitectura, salud, movilidad, clima, grupo étnico, grupo social, lengua y medidas (FIG. 5).

Fig. 5. Ejemplo de una de las Relaciones Geográficas anotadas.

Para conectar la información histórica de la colección con las geografías que ésta menciona, fue necesario crear un *gazetteer* o nomenclátor histórico-geográfico. Este se trata de un diccionario o índice exhaustivo de topónimos con sus coordenadas y otra información geográfica del Siglo XVI. El *DECM gazetteer* contiene información sobre la ubicación, los límites y las características físicas de las entidades geográficas registradas en las Relaciones Geográficas (FIG. 6). Registra casi 15,000 topónimos y variantes lingüísticas, y va acompañado de otras 49 capas de información histórica relevante y tablas con datos sobre lenguas, mapas, repositorios, etc. Está disponible en formatos GIS y JSON. Otros productos y artículos elaborados por el proyecto pueden consultarse en nuestro sitio web (https://www.lancaster.ac.uk/digging-ecm/corpus/) y en el repositorio de Github (https://github.com/patymurrieta/Digging-into-Early-Colonial-Mexico/).

Fig. 6. DECM Gazetteer. El diccionario geográfico digital mostrando los topónimos mencionados en toda la colección de las Relaciones Geográficas de Siglo XVI.

Otro proyecto actualmente en curso es "Unlocking the Colonial Archive: Harnessing Artificial Intelligence for Spanish and Indigenous historical collections' (https://unlockingarchives.com/). Se trata de una colaboración entre humanistas digitales, historiadores, arqueólogos, informáticos, ingenieros y lingüistas, así como varias instituciones del Reino Unido, México y Estados Unidos. El proyecto es financiado por el AHRC-UK y NEH-US. En colaboración con la LLILAS Benson Latin American Studies and Collections de la Universidad de Texas en Austin, el Fondo Real de Cholula, y el Archivo General de la Nación, el proyecto tiene tres objetivos. El primero es aprovechar el reconocimiento de textos manuscritos para agilizar la transcripción de miles de documentos históricos e impresos de estas colecciones. El segundo es extraer información de estos textos combinando PLN y Datos Abiertos Vinculados como ya se ha explicado arriba. El tercero es utilizar las técnicas de visión por computadora para facilitar la identificación, extracción, búsqueda y análisis automatizados de características pictóricas en documentos indígenas y coloniales. En este proyecto estamos desarrollando diferentes técnicas en estas áreas y, para finales de 2025, esperamos disponer de modelos de aprendizaje automático que ayuden en la transcripción automatizada de 5 tipos diferentes de caligrafía: Gótica, Cortesana, Itálica cursiva, Procesal y Procesal Encadenada.

También esperamos contar con un modelo de PLN para realizar minería de textos a partir de fuentes en español y náhuatl, así como una serie de enfoques para convertir datos de archivos en colecciones de Datos Abiertos Vinculados, y al menos un modelo de visión por computadora para la identificación automatizada y el análisis de características en pinturas e imágenes. Esto tiene el fin de conectar Information como las colecciones digitalizadas de fuentes Mesoamericanas alrededor del mundo con la red semántica como se explicó anteriormente.

Algunas reflexiones finales: Una postura crítica ante el uso de las herramientas computacionales y el camino hacia la Inteligencia Artificial Decolonial

Espero que las pocas líneas anteriores hayan demostrado que estamos viviendo tiempos apasionantes en los cuales se están abriendo nuevos caminos en la investigación de la historia novohispana. A medida que el panorama digital evoluciona y los enfoques y tecnologías científicas se entrelazan con las humanidades, surgen métodos y herramientas de formas inéditas y novedosas. Pero esto también va acompañado de los retos que plantean estos enfoques y materiales. Las tecnologías y los programas computacionales nunca son neutrales, apolíticos, o inocentes. Las herramientas y métodos computacionales, incluidos todos los de la Inteligencia Artificial, no se crean en un vacío intelectual, cultural o social. Esto significa que cada herramienta, infraestructura y tecnología nace en un contexto y una cosmovisión específica, y debemos ser conscientes de que ésta puede no ser compartida por todos, donde existe el riesgo de reproducir estructuras no deseadas. Por ejemplo, hemos señalado en otro lugar (Murrieta-Flores, Favila Vázquez y Flores Morán, 2022), cómo tecnologías como los SIG son producto de una concepción cartesiana, euclidiana y occidental del espacio que no es común a todas las culturas, y que al utilizar esta tecnología para analizar procesos históricos imponemos una lente y hacemos suposiciones de las que debemos ser muy conscientes, particularmente cuando investigamos aquellas sociedades que no pertenecen a dicha tradición, incluyendo el mundo mesoamericano y el colonial. Además, casi todas las tecnologías digitales y los desarrollos de software han tenido lugar en el Norte Global, y han sido creados mayoritariamente por hombres blancos que suelen provenir de clases sociales privilegiadas. Esto es inequívocamente un producto de la historia geopolítica, patriarcal y hegemónica del mundo. Los conjuntos de datos que se utilizan para entrenar algoritmos de aprendizaje automático también proceden principalmente del Norte Global, están en inglés o, por lo general en lenguas europeas modernas, y ya sean textos o imágenes, son predominantemente representativos de una tradición occidental (Great Learning, 2020; Towards AI, 2022). En realidad, tanto en los datos como en los algoritmos, hay poca o ninguna representación de comunidades indígenas o subalternas, mujeres, u otros grupos del Sur Global. El resultado de esto son algoritmos y herramientas racistas, misóginas y clasistas, y a pesar de que los prejuicios son bien conocidos (Irani et al., 2010; Mohamed et al., 2020; Mehrabi et al., 2021), solo ahora están empezando a surgir algunas iniciativas para cambiar esta situación. Como dice Gabor Maté, "Puede que no seamos responsables del mundo que creó nuestras mentes, pero podemos responsabilizarnos de la mente con la que creamos nuestro mundo" (2007). Así, con la colaboración de importantes instituciones culturales, comunidades indígenas y locales, investigadores, industria, y empresas sociales de 22 países de todo el mundo, especialmente del Sur Global, estamos trabajando para establecer un centro de investigación sobre Inteligencia Artificial decolonial.

Hoy en día, la norma son las infraestructuras y las herramientas que suelen acomodarse a una visión del mundo específica (es decir la occidental), y los datos sesgados perpetúan y amplifican las

desigualdades sociales pasadas y presentes. Esto tiene que cambiar. Hay que recordar que nuestras actuales situaciones económicas, sociales y políticas surgieron y están profundamente arraigadas en el otrora sistema mundial colonial en todo su Occidentalismo, dejando poco o ningún espacio para lo que Gloria Anzaldúa llamó "pensamiento fronterizo" (Anzaldúa, 2012). Considero entonces que es necesario hacer un llamado en el desarrollo de las HD en América Latina, a transformar las rígidas fronteras epistémicas y territoriales establecidas en el proceso de construcción del sistema del mundo moderno y controladas aún por la colonialidad del poder (Mignolo, 2012: 12). También propongo reutilizar para el desarrollo de enfoques computacionales en historia novohispana, lo que Foucault llamó hace cuarenta y dos años "la insurrección de los saberes subyugados" (1980), creando no sólo tecnologías y conjuntos de datos que se ajusten a nuestras propias cosmovisiones en el Sur Global, sino también invirtiendo a través de la colaboración, las jerarquías y la subalternización del conocimiento que fueron impuestas por la colonialidad, ya sea que esto se refiera al imaginario colonial sobre el indígena en el contexto histórico, al masculino sobre todos los demás en el contexto de género, o al científico sobre el humanístico en el contexto del saber. Casi veinte años después de mis primeras andanzas en el AGN y mi incursión a la informática como campo, estoy convencida de dos cosas. Primero, que debido al encuentro altamente interdisciplinario entre las ciencias y las humanidades emergiendo ahora en Latinoamérica, las Humanidades Digitales son la disciplina más apta para liderar esta revolución, y segundo, que campos como la historia novohispana y la arqueología histórica que ahondan en la investigación crítica y evidencian los profundos desequilibrios de poder y las dinámicas coloniales, son cruciales en este empeño.

Agradecimientos

La investigación mencionada en este artículo es producto de los proyectos ‘Digging into Early Colonial Mexico’ UKRI-ESRC, y ‘Unlocking the Colonial Archive: Harnessing Artificial Intelligence for Spanish American Historical Collections’ patrocinado por UKRI-AHRC, Grant AH/V009559/1.

Bibliografía

Anzaldúa, Gloria. 2012. *Borderlands / La Frontera: The New Mestiza*. Edited by Norma Cantú and Aída Hurtado. Revised edition. San Francisco: Aunt Lute Books.

Chowdhary, K. R. 2020. ‘Natural Language Processing’. In *Fundamentals of Artificial Intelligence*, edited by K.R. Chowdhary, 603–49. New Delhi: Springer India. https://doi.org/10.1007/978-81-322-3972-7_19.

Cooper, David, and Ian N Gregory. 2011. ‘Mapping the English Lake District: A Literary GIS: Mapping the English Lake District’. *Transactions of the Institute of British Geographers* 36 (1): 89–108.

Donaldson, Christopher, Ian Gregory, and Patricia Murrieta-Flores. 2015. ‘Mapping “Wordsworthshire”: A GIS Study of Literary Tourism in Victorian Lakeland’. *Journal of Victorian Culture* 20 (3): 287–307.

Fernández-Aceves, Hervin. 2017. ‘Social Network Analysis and Narrative Structures: Measuring Communication and Influence in a Medieval Source for the Kingdom of Sicily’. *Intersticios Sociales*, no. 14 (August): 125–53. https://doi.org/10.55555/IS.14.148.

Foucault, Michael. 1980. *Power/Knowledge: Selected Interviews and Other Writings, 1972-1977*. London: Vintage. https://www.amazon.de/-/en/Michel-Foucault/dp/039473954X.

Great Learning. 2020. 'Top 5 Sources For Analytics and Machine Learning Datasets'. Great Learning Blog: Free Resources What Matters to Shape Your Career! 11 May 2020. https://www.mygreatlearning.com/blog/sources-for-analytics-and-machine-learning-datasets/.

Hernández Pons, Elsa. 2000. 'Arqueología histórica en México: su situación actual'. *Arqueología - Instituto Nacional de Antropología e Historia*, no. 23: 103–26.

Hutchinson, Tim. 2020. 'Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing'. *Records Management Journal* 30 (2): 155–74. https://doi.org/10.1108/RMJ-09-2019-0055.

Irani, Lilly, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. 'Postcolonial Computing: A Lens on Design and Development'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1311–20. CHI '10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1753326.1753522.

Jordan, M. I., and T. M. Mitchell. 2015. 'Machine Learning: Trends, Perspectives, and Prospects'. *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Lausch, Angela, Andreas Schmidt, and Lutz Tischendorf. 2015. 'Data Mining and Linked Open Data – New Perspectives for Data Analysis in Environmental Research'. *Ecological Modelling*, Use of ecological indicators in models, 295 (January): 5–17. https://doi.org/10.1016/j.ecolmodel.2014.09.018.

Maté, G. 2007. *In the Realm of Hungry Ghosts: Close Encounters with Addiction*. London: Random House UK.

McEnery, Tony. 2019. *Corpus Linguistics*. Edinburgh University Press.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. 'A Survey on Bias and Fairness in Machine Learning'. *ACM Computing Surveys* 54 (6): 115:1-115:35. https://doi.org/10.1145/3457607.

Mignolo, Walter. 2012. *Local Histories/Global Designs Coloniality, Subaltern Knowledges, and Border Thinking*. Princeton Studies in Culture/Power/History. Woodstock: Princeton University Press.

Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. 'Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence'. *Philosophy & Technology* 33 (4): 659–84. https://doi.org/10.1007/s13347-020-00405-8.

Murrieta-Flores, Patricia, Alistair Baron, Ian Gregory, Andrew Hardie, and Paul Rayson. 2015. 'Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century: Automatically Analyzing Large Historical Texts in a GIS Environment'. *Transactions in GIS* 19 (2): 296–320.

Murrieta-Flores, Patricia, Mariana Favila-Vázquez, and Aban Flores-Morán. 2022. 'Indigenous Deep Mapping: A Conceptual and Representational Analysis of Space in Mesoamerica and New Spain'. In *Making Deep Maps*, edited by David J. Bodenhamer, John Corrigan, and Trevor M. Harris, 1st ed., 78–111. London: Routledge. https://doi.org/10.4324/9780367743840-6.

Murrieta-Flores, Patricia, and Ian Gregory. 2015. 'Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology'. *Open Archaeology* 1 (1): 166–75. https://doi.org/10.1515/opar-2015-0010.

Murrieta-Flores, Patricia, Diego Jiménez-Badillo, and Bruno Martins. 2022. 'Digital Resources: Artificial Intelligence, Computational Approaches, and Geographical Text Analysis to Investigate Early Colonial Mexico'. In *Oxford Research Encyclopedia of Latin American History*, by Patricia Murrieta-Flores, Diego Jiménez-Badillo, and Bruno Martins. Oxford University Press. https://doi.org/10.1093/acrefore/9780199366439.013.977.

Pereda, Javier. 2019. 'A TUI to Explore Cultural Heritage Repositories on the Web'. In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 259–67. TEI '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3294109.3301000.

Shapiro, Linda G., and George C. Stockman. 2001. *Computer Vision*. 1st edition. Upper Saddle River, NJ: Pearson.

Sporleder, Caroline. 2010. 'Natural Language Processing for Cultural Heritage Domains'. *Language and Linguistics Compass* 4 (9): 750–68. https://doi.org/10.1111/j.1749-818X.2010.00230.x.

Towards AI. 2022. 'Best Public Datasets for Machine Learning and Data Science'. Medium. 30 September 2022. https://pub.towardsai.net/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f.

Voulodimos, Athanasios, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. 'Deep Learning for Computer Vision: A Brief Review'. *Computational Intelligence and Neuroscience* 2018 (February): e7068349. https://doi.org/10.1155/2018/7068349.

Won, Miguel, Murrieta-Flores, Patricia, and Martins, Bruno. 2018. 'Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora'. *Frontiers in Digital Humanities* 5.

Yu, Liyang. 2011. 'Linked Open Data'. In *A Developer's Guide to the Semantic Web*, edited by Liyang Yu, 409–66. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-15970-1_11.

Zhou, Zhi-Hua. 2021. *Machine Learning*. Springer Nature.