



La realidad que emerge tras la

avalancha de datos

Las humanidades digitales se propagan por diversas disciplinas, influyen en la formación de investigadores e inspiran políticas públicas

Fabrizio Marques

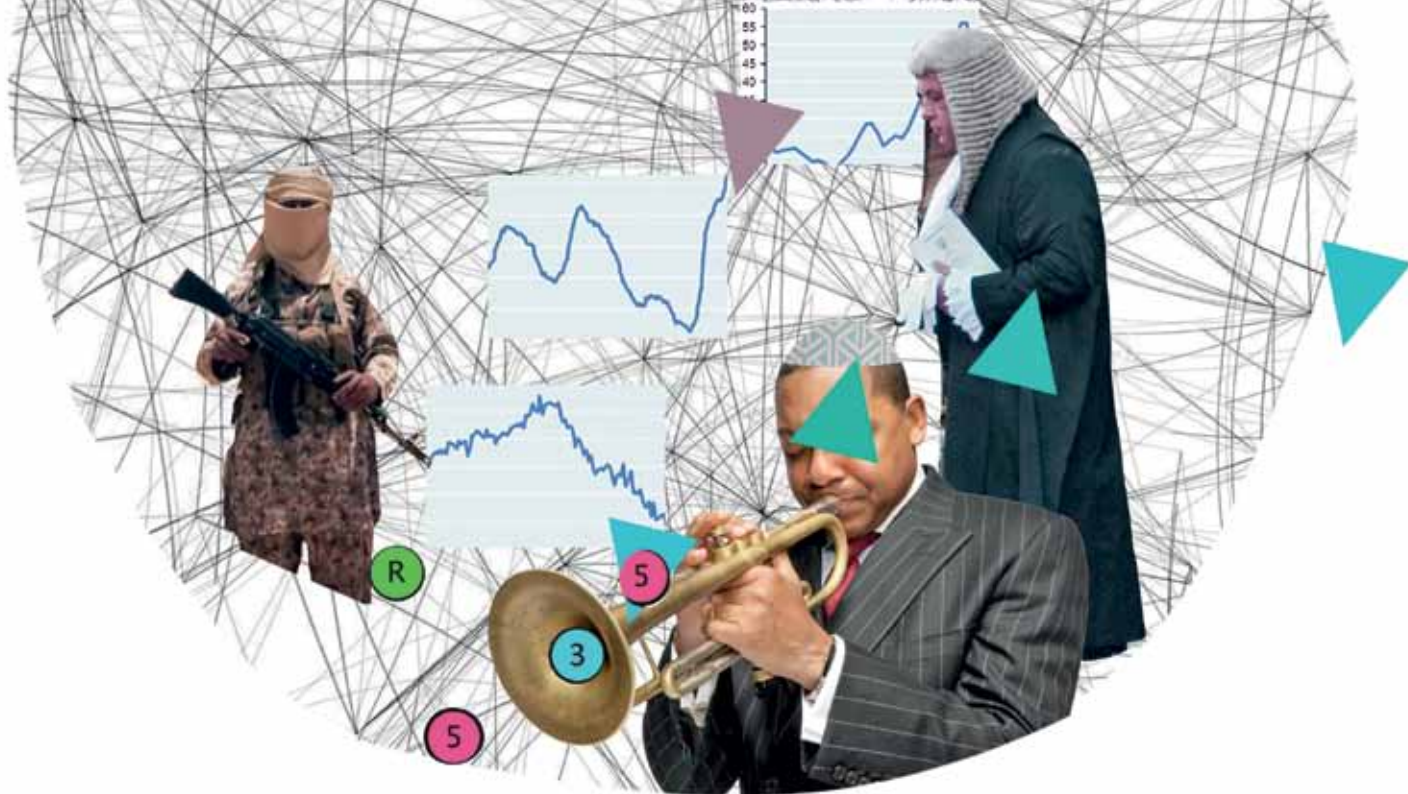
PUBLICADO EN MAYO DE 2017

Las computadoras son herramientas de trabajo para los investigadores de todas las áreas del conocimiento, pero en el caso de la comunidad de las ciencias humanas y sociales, la digitalización de los archivos artísticos e históricos, así como la oferta de gigantes bancos de datos con informaciones económicas y sociales, han abierto nuevos frentes de observación de fenómenos y de análisis de tendencias. Se ha producido un natural acercamiento con los científicos de la computación, cuyas investigaciones en *big data* multiplicaron las formas de organizar y analizar informaciones, dando origen a un campo interdisciplinario: las humanidades digitales. “Este término fue acuñado para definir a la investigación científica que incorpora la tecnología informática a los estudios en humanidades, pero también a aquella que se vale de las humanidades para estudiar la tecnología digital y su influencia en la sociedad y en la cultura”, explica Brett Boley, director de la Oficina de Humanidades Digitales de la National Endowment for the Humanities (NEH), una agencia de fomento del gobierno estadounidense. No se trata, según Boley, de una nueva área del conocimiento,

sino de una gama de actividades que puede abarcar el uso de fotografías aéreas por arqueólogos para escanear sitios, el desarrollo de técnicas de análisis de datos que ayudan a los lingüistas en el estudio de periódicos antiguos y el estudio de la ética de la tecnología por filósofos, entre otros ejemplos.

Uno de los proyectos financiados por la NEH en humanidades digitales recuperó los diarios de campo del explorador británico David Livingstone (1813-1873). Relatos de su viaje a África central en 1871 fueron redactados en periódicos antiguos a falta de papel disponible. Con el paso del tiempo, la tinta se desvaneció y quedaron ilegibles los textos en donde Livingstone registró sus impresiones acerca de la dinámica del comercio de esclavos, entre otras anotaciones. Entre 2013 y 2017, un grupo de investigadores de humanidades y ciencias de la computación de Estados Unidos y del Reino Unido logró recuperar esos escritos valiéndose de técnicas de fotografía de imagen espectral, capaces de recuperar información que no es visible al ojo humano.

Otro ejemplo fue la colaboración de historiadores de distintas partes del mundo para organizar registros de alrededor de 36 mil viajes de buques negreros entre



1514 y 1866, que transportaron a más de 12 millones de esclavos desde África. Esta labor, iniciada en los años 1990 por el historiador estadounidense David Eltis, de la Universidad Emory, resultó en el Banco de Datos sobre el Comercio de Esclavos, disponible en internet desde 2007 en el portal slavevoyages.org. El estudio de esos datos, que recopila registros en varios idiomas e incluye todo el movimiento en los puertos por donde pasaron los barcos, indicó a los historiadores nuevas dimensiones al respecto del modo en que los africanos sintieron y resistieron la deportación y la esclavización, revelando nuevas conexiones transatlánticas en el marco del comercio de esclavos.

En 1999 apareció un primer estudio en formato de CD-ROM, pero el esfuerzo conjunto para obtener datos sobre los viajes consiguió trazar posteriormente un retrato más amplio del comercio de esclavos. En una primera etapa, se calcula que Brasil habría recibido alrededor de 3,6 millones de esclavos, pero los documentos revelaron que ese contingente llegó a ser de 5 millones, sobre un total de 10,7 millones de africanos deportados hacia América. Esa iniciativa produjo diversos impactos en las investigaciones sobre la esclavitud, dice Manolo Florentino, docente de la Universidad Federal de Río de Janeiro (UFRJ) y responsable del brazo brasileño del proyecto. El más destacable fue que reemplazó estimaciones por datos concretos, obtenidos de fuentes primarias. Otro dato fue que reveló la preponderancia brasileña en el comercio de esclavos. “Buena parte de los documentos obtenidos en el marco del proyecto están escritos en portugués, una especie de lengua franca del tráfico negrero”, comenta Florentino, quien en años recientes se dedicó a traducir

El banco de datos sobre el tráfico negrero reveló la preponderancia de Brasil en el marco del comercio de esclavos

todo el sitio web en portugués. La colección de datos sobre la deportación y la esclavización de los africanos permite ahora, según el historiador, alimentar un frente de investigación menos explorado, constituido por las trayectorias recorridas por los esclavos por el interior de Brasil, luego de su arribo a los puertos.

UNA DIVERSIDAD DE PROYECTOS

Los resultados de una reciente llamada a la presentación de proyectos evidenciaron la diversidad de las humanidades digitales. La cuarta edición de la convocatoria denominada Digging Into Data Challenge registró 108 propuestas enviadas por equipos interdisciplinarios de 11 países, de las cuales se aprobaron 14. Esta iniciativa forma parte de la Plataforma Transatlántica (T-AP), un trabajo conjunto en ciencias humanas y sociales que agrupa a 16 agencias de fomento de Europa y América, entre las cuales figura la FAPESP. “Tuvimos un aumento notorio de países participantes, que en llamados anteriores eran tan sólo cuatro. Esto marca una gran diferencia, con el surgimiento de nuevas co-

laboraciones”, dice Brett Bobley, mentor del programa Digging Into Data en 2008. Los proyectos aprobados están distribuidos por disciplinas, entre las cuales figuran musicología, lingüística, historia, ciencias políticas y economía, y recibirán inversiones que, en total, suman 9,2 millones de dólares, el equivalente a 29 millones de reales. Una de las propuestas contempladas reúne a investigadores de Estados Unidos, Alemania y Holanda, y estará abocada al estudio de tres bancos de datos que congregan registros escritos y orales sobre el folklore en varios puntos de Europa. La meta consiste en la identificación de patrones que se repitan a lo largo del tiempo en lugares diferentes y ayuden a develar cuáles eran las creencias comunes en el pasado, basándose en las historias contadas y en la propagación de leyendas y casos sobrenaturales.

Otro de los proyectos, liderado por economistas y científicos de la computación de Estados Unidos, Canadá y Holanda, apunta a cruzar informaciones sobre la variación de los precios de productos vendidos por internet en todo el mundo, recopilados permanentemente por el proyecto Billion Prices, del Massachusetts Institute of Technology (MIT), con datos económicos para elaborar estudios sobre la inflación, el poder adquisitivo y el estándar de vida en varios países. También hay una iniciativa en cuyo marco se analizarán 70 años de cobertura de la prensa sobre ataques terroristas, en busca de modelos de lo que sería un abordaje responsable del problema, e incluso otra que investigará las estructuras melódicas en grabaciones de jazz, intentando asociarlas a la evolución del contexto histórico y social en el cual surgieron esas canciones.

En la selección de los 14 proyectos contemplados intervinieron más de 200 expertos que evaluaron las 108 propuestas. “La diversidad de los problemas abordados revela un gran potencial para el desarrollo de las humanidades digitales en Brasil”, comenta Claudia Bauzer Medeiros, docente del Instituto de Computación de la Universidad de Campinas (Unicamp) y representante de la FAPESP en la T-AP, quien participó en todo el proceso, desde la preparación del pliego hasta la selección de los proyectos. “Este campo aún está poco explorado en el país porque aquí todavía no hay tanta colaboración entre investigadores de las ciencias humanas y sociales y los de la computación. Poco a poco ellos van percibiendo que tal interacción es algo factible. No es necesario que el investigador en ciencias humanas y sociales sea un entendido en computación para trabajar en dicha área, pero es preciso que colabore con expertos en los aspectos informáticos”, dice la investigadora, quien coordina el Programa de Investigación en e-Science de la FAPESP.

Londres contra el crimen

Herramientas que estudian datos sobre 197 mil juicios

Registros acerca de 197 mil juicios realizados entre 1674 y 1913 en el Tribunal Penal Central de Londres, más conocido como Old Bailey, que es el nombre de la calle en donde está situada la corte, fueron puestos a disposición para la consulta en internet a partir de 2003 en la dirección electrónica oldbaileyonline.org. El reto de identificar fenómenos y tendencias en medio de un volumen de información que llega a 127 millones de palabras movilizó a investigadores del Reino Unido y de Estados Unidos, que desarrollaron formas de explorar datos textuales bastante más sofisticados que los de la búsqueda disponible en el repositorio.

El proyecto Data Mining with Criminal Intent, financiado en 2009 en el marco del primer pliego del llamado a la presentación de propuestas denominado Digging Into Data, escudriñó en los registros de Old Bailey con la ayuda de una combinación de herramientas digitales. Una de ellas es la denominada Zotero, que permite recabar y ordenar información, y otra, un portal denominado TAPoR, que ayuda a los usuarios en el análisis de textos utilizando diferente *software*. Esa estrategia permitió llegar a resultados singulares. Por ejemplo, se pudo comprobar que la palabra “veneno” estaba asociada más frecuentemente con “café” que con “comida”, evidenciando la forma en que los londinenses eran asesinados por envenenamiento.

Del mismo modo, pudo observarse que las penas para los bígamos atenuaron su severidad con el transcurso del siglo XIX. Según Stephen Ramsay, profesor de inglés de la Universidad de Nebraska-Lincoln, uno de los líderes de la iniciativa, el aporte del proyecto no se limita a la obtención de evidencias históricas que anteriormente no eran percibidas. “Las historias de Old Bailey expresan las motivaciones más densas de la condición humana, tales como la venganza, la deshonra y la pérdida, que son la materia prima de las humanidades”, dijo, según publica el periódico *The Chronicle of Higher Education*.



Uno de los proyectos seleccionados en el marco del Digging Into Data Challenge cuenta con la participación de brasileños. Se trata de una colaboración entre investigadores de Francia, Argentina y Brasil que apunta a estudiar cómo se propagan las opiniones en la sociedad y cómo ese proceso sufrió transformaciones con el avance de la tecnología de la información. En el estudio se analizarán dos bancos de datos para mapear la construcción de redes de relaciones entre grupos de individuos, donde esas conexiones estarán representadas en estructuras visuales, los grafos. En uno de los bancos de archivos, el del periódico *The New York Times*, el objetivo será analizar los artículos sobre Brasil publicados a lo largo de 70 años para mapear las relaciones entre grupos de individuos y entidades mencionados en esos textos referidos al país. “La intención es comprender de dónde provenían y cómo se relacionaban las ideas y opiniones reproducidas en los textos, principalmente aquéllas que versaban sobre temas políticos y económicos, y cómo evolucionó eso con el tiempo, así como verificar la posible influencia de las noticias publicadas en el periódico por corresponsales extranjeros en la formación de la opinión pública en el país”, explica la investigadora Maria Eunice Quilici Gonzalez, líder del grupo brasileño que participa



en el proyecto y docente del Departamento de Filosofía de la Facultad de Filosofía y Ciencias de la Universidade Estadual Paulista (Unesp), en su campus de Marília.

El segundo banco de datos es una colección de publicaciones sobre procesos electorales en la red social Twitter. La idea es mostrar cómo se constituyen y se consolidan las opiniones en un ambiente virtual. “Pretendemos analizar la dinámica de la propagación de las opiniones en las redes sociales. Cuanto más frecuentes son las relaciones, más densos se tornan los nodos de las redes representadas en los grafos. La tendencia

Cómo se urbanizó São Paulo

Una plataforma aglutinará datos georreferenciados sobre la transformación de la capital paulista entre 1870 y 1940

La urbanización de São Paulo ocurrió con mayor velocidad que la de otras metrópolis, saltando de tan sólo 30 mil habitantes en 1870 a 1 millón de habitantes en 1940. El estudio de las transformaciones que sufrió la ciudad durante ese período contará con el respaldo de una plataforma con informaciones georreferenciadas que serán provistas por innumerables fuentes: tesis, informes o mapas. Cualquier investigador que disponga de datos y pueda relacionarlos con una dirección electrónica de la capital paulista está invitado a incluirlos en la plataforma Pauliceia 2.0, cuyo proyecto fue presentado a potenciales usuarios el 4 de abril, en busca de sugerencias.

El proyecto, que congrega a investigadores de la Universidad Federal de São Paulo (Unifesp), del Instituto Nacional de Investigaciones

Espaciales (Inpe), del Archivo Público del Estado de São Paulo y de la Emory University, de Estados Unidos, cuenta con financiación del Programa FAPESP de Investigación en e-Science. “Quien haya estudiado los hoteles de São Paulo podrá agregar a las correspondientes direcciones información sobre cada uno de ellos. Aquéllos que hayan estudiado los delitos cometidos en la ciudad, también. Cualquier información que pueda ubicarse en forma espacial puede alimentar la plataforma”, dice el historiador Luis Ferla, docente de la Unifesp que es quien coordina el proyecto.

Dentro del proyecto hay un equipo abocado al desarrollo de un banco de datos con la numeración de las edificaciones de la época, para asegurar que la localización de las informaciones sea fidedigna. “Es un trabajo tan complejo que en primera instancia se



lo está testeando en un área piloto en el centro de São Paulo”, explica Ferla. Para julio de 2018 estará disponible una versión preliminar de la plataforma para pruebas. “Quienes deseen estudiar ese período encontrarán mucho material en la plataforma para elaborar sus propias reflexiones. El proyecto propone la realización de una curaduría del conocimiento sobre la urbanización de la ciudad”. Más información estará disponible en la siguiente dirección electrónica: unifesp.br/himaco.

La ciudad en la década de 1940, cuando llegó a su primer millón de habitantes



indica que cobran centralidad e inhiben el crecimiento de otros nodos, revelando la trayectoria de la formación de una opinión”, informa Quilici Gonzalez. Uno de los intereses reside en el estudio de la formación de ámbitos de polarización política en las redes sociales. “Grupos que antiguamente se encontraban aislados logran fortalecer sus opiniones ganado adeptos, alimentándose de las comunicaciones en las redes sociales. Eso es lo que sucedió recientemente, por ejemplo, con los grupos a favor o en contra del *impeachment* [el proceso de destitución o juicio político] en Brasil”. Más allá de sus objetivos específicos, el proyecto también tiene ambiciones más generales, entre las cuales figura la de evaluar la posibilidad de crear modelos para el estudio de las actividades sociales e investigar las posibles consecuencias éticas del uso del análisis de macrodatos en procesos de autoorganización social, emergentes de la interacción espontánea entre varios actores sociales, sin el liderazgo o la interferencia de un centro organizador.

El proyecto se realizará en colaboración con investigadores de las universidades de Cergy-Pontoise, en Francia, y de Buenos Aires (UBA) en Argentina. El equipo es crítico de la tesis que sostiene que pueden moldearse comportamientos o encauzar la formación de la opinión manipulando las tendencias obtenidas solamente mediante el análisis de *big data*. “Resulta exagerado afirmar que la elección de Donald Trump como presidente estadounidense o la salida de la Unión Europea que votaron los británicos se deba exclusivamente a las respectivas campañas que utilizaron los servicios de una empresa de *marketing* político –Cambridge Analytica– que se habría basado en datos y herramientas de las redes sociales para manipular los temores y anhelos de los electores”, dice Quilici Gonzalez. “El estudio del *big data* puede apuntar tendencias, pero está lejos de explicar la naturaleza humana. Su empleo sólo será eficiente de estar acompañado del estudio de las disposiciones de ciertos grupos que, en los casos de Estados Unidos y del Reino Unido, estaban relacionadas con la preeminencia de cierto nacionalismo que reniega del multiculturalismo”.

Quilici Gonzales se graduó en física, con máster en filosofía y doctorado en lingüística y ciencia cognitiva. Colaborará con el proyecto, con la ayuda de un equipo de científicos brasileños, con reflexiones relativas a la ética implícita en las intervenciones de los individuos en las redes sociales. “El concepto de privacidad, por ejemplo, es algo que está cambiando. Algunas de las nociones de privacidad de mi generación no se aplican a los sujetos en las redes sociales, que exponen sistemáticamente detalles personales. También está el problema de aquellos individuos

que crean perfiles falsos, alterando sus características personales, su situación socioeconómica e incluso su género para interactuar virtualmente con otros individuos”, dice. Según la investigadora, si en su intimidad a menudo una persona tiene que mantener una identidad que no le agrada, en las redes sociales puede realizar sus fantasías sin las supuestas presiones familiares. “La identidad es ficticia, pero la información que el individuo proporciona puede ser real, en cierto sentido. Por medio de ella puede crear una relación con compañeros virtuales, algo que antiguamente no existía”. Para estudiar situaciones de este tipo, el grupo brasileño razonará sobre el modo en que el análisis de macrodatos puede ayudar en la comprensión de nuevos modelos de conducta y de la dinámica de la formación de la opinión colectiva.

TEMAS Y AVANCES

El programa de la próxima edición del simposio Digital Humanities, que congregará en el mes de agosto a alrededor de mil investigadores de varias nacionalidades en la ciudad de Montreal, en Canadá, brinda una dimensión de los temas y de los avances que establecieron puentes entre los científicos de la computación y los profesionales de las ciencias humanas y sociales. Habrá *workshops* que tratarán sobre temas tales como la aplicación en investigaciones de humanidades de herramientas de visualización computarizada, un concepto que se utiliza principalmente en el campo de la robótica y por medio del cual, los sistemas artificiales son capaces de extraer informaciones de imágenes simulando el funcionamiento de la visión humana. O habrá debates sobre problemas éticos y legales relacionados con el uso de datos digitalizados que pueden exponer la privacidad de los individuos. En el marco del encuentro en Montreal serán homenajeados los responsables del proyecto Text Encoding Initiative (TEI), un consorcio que desde hace 30 años desarrolla y mantiene un modelo para la codificación de textos en formato digital que los torna legibles por máquinas, y que impulsó investigaciones en ciencias humanas, principalmente en el área de la lingüística. “En los últimos 15 años se produjo un cambio cualitativo en el volumen de los datos textuales disponibles, algo que modificó radicalmente las posibilidades de investigación”, afirma Karina van Dalen-Oskam, presidente de la Alianza de Organizaciones de Humanidades Digitales (ADHO), la entidad organizadora de la conferencia. Docente de estudios literarios computacionales de la Universidad de Ámsterdam, en Holanda, Dale-Oskam resalta el progreso de los nuevos abordajes para la investigación en literatura, como el concepto de lectura distante, que analiza grandes volúmenes de datos relacionados no sólo con la obra en estudio, sino con la

Un corpus histórico de la lengua portuguesa

Un banco de datos con 3,3 millones de palabras reúne anotaciones sobre textos de distintas épocas

En algunas áreas de las humanidades, la cooperación con los científicos de la computación se produjo en forma más natural que en otras. Un ejemplo de ello lo constituyen los estudios sobre las transformaciones en el uso de la lengua. Charlotte Galves, docente del Instituto de Estudios del Lenguaje de la Universidad de Campinas (IEL-Unicamp), suele decir que se dedicaba a las humanidades digitales mucho antes de saber que existía esa denominación. En 1998, la investigadora comenzó a recopilar textos de los siglos XVI al XIX para elaborar un corpus histórico de la lengua portuguesa, un banco de textos con anotaciones morfosintácticas de palabras y oraciones que ha servido de base para la realización de una serie de estudios sobre la historia del idioma portugués en Portugal y en Brasil. “Se está pudiendo observar cómo se transformó el idioma en el curso de los siglos, particularmente en Brasil, donde se viene escindiendo del portugués europeo al influjo del contacto con otras lenguas, pese a que volvió a sufrir su influencia durante la segunda mitad del siglo XIX”, comenta Galves.

El banco de datos fue creciendo y actualmente cuenta con 3,3 millones de palabras provenientes de 76 textos originales. El archivo, bautizado con el nombre de Corpus Tycho Brahe, en referencia al astrónomo danés del siglo XVI que se propuso catalogar el desplazamiento de los planetas, tuvo sus primeras herramientas para etiquetar palabras desarrolladas por el científico de la computación Marcelo Finger, docente del Instituto de



Los escritos de sacerdote Antônio Vieira (1608-1697) forman parte del archivo

Matemática y Estadística de la USP. La evolución fue lenta. Las correcciones de las anotaciones automáticas las fue haciendo Galves en forma personal, con la ayuda de posdoctores y dirigidos. “Aprendí mucho acerca del *big data*, pero no podría prescindir de la ayuda de los científicos de la computación”, dice Galves. El paso siguiente consiste en tornar integralmente accesible el banco de datos vía internet –actualmente puede realizarse la descarga del archivo en la siguiente dirección electrónica: tycho.iel.unicamp.br/corpus-, pero no búsquedas *online*.

El mismo modelo del portugués histórico ahora lo están utilizando Galves y Filomena Sandalo, también docente de la Unicamp, para el estudio de una lengua indígena, el idioma kadiwéu, hablado por una etnia que habita en el estado de Mato Grosso. Se recopilaron relatos orales de los aborígenes y se los está convirtiendo en textos escritos con anotaciones. “La idea es generar un corpus lingüístico de otros idiomas dentro de la misma plataforma, utilizando las mismas herramientas”, explica Galves.

totalidad del contexto histórico en que la misma fue elaborada, o el campo de la estilometría, que posibilita detectar la autoría de textos apócrifos. “Tales abordajes permiten saber más sobre el desarrollo de géneros literarios e incluso sobre aquellos factores que hacen que un texto se torne o no un *best seller*”, dice.

El crecimiento de este campo interdisciplinario coexiste con críticas de que las humanidades digitales producirían más titulares que avances sólidos del conocimiento y también que rivalizan con los campos tradicionales de las humanidades en el reparto de la financiación a la investigación. En un artículo publicado en el periódico *The New York Times* en 2015, Armand Marie Leroi, docente de biología evolutiva del Imperial College de Londres, en el Reino Unido, puso en duda la capacidad de las humanidades digitales para producir análisis innovadores en la literatura. A su juicio, la posibilidad de convertir arte en datos torna posible hallar nuevos significados para una obra mediante el uso de nuevos algoritmos. “Pero será necesario crear un algoritmo muy preciso para que sea capaz de detectar la ironía en la obra de Jane Austen”, escribió. “La verdad de la crítica artística no es del mismo tipo que la verdad científica”.

Los investigadores del área argumentan que las humanidades digitales ofrecen tan sólo una extensión de los métodos y habilidades tradicionales sin ambición de sustituirlos. El libro *Digital Humanities* (MIT Press, 2012), redactado por un conjunto de autores, sostiene en su primer capítulo que las humanidades digitales “no obliteran las ideas del pasado, sino que suplementan el compromiso de las humanidades con la interpretación académica, la investigación informada, el argumento estructurado y el diálogo entre las comunidades que la practican”.

El politólogo Eduardo Marques, docente de la Facultad de Filosofía, Letras y Ciencias Humanas de la Universidad de São Paulo (FFLCH-USP), subraya que los abordajes de la ciencia de la computación y de las ciencias humanas y sociales dentro de las humanidades digitales tienen orígenes diferentes. “Hubo un encuentro de dos movimientos. Uno proveniente de las ciencias duras, con el desarrollo de herramientas de búsqueda de datos que posibilitaron la producción de información sobre el mundo social y la generación de nuevos campos empíricos. En tanto, en las ciencias humanas, se partió del uso preestablecido de herramientas estadísticas para el estudio de fenómenos sociales”, explica. Como las lógicas son distintas, resulta complicado unificarlas, dice Marques. “Mientras que los científicos de la computación buscan modelos en los grandes volúmenes de datos para enunciar preguntas de investigación, los científicos



Las carreras y materias sobre análisis cuantitativo y ética en el uso de datos van ganando espacio

sociales parten de supuestos teóricos y utilizan herramientas digitales para cotejar su validez. El diálogo es rico, pero es difícil unificar las diferentes formas de encarar el asunto”.

Dicho diálogo está influyendo en la formación de los investigadores. En el caso de las ciencias humanas y sociales, las carreras y las materias sobre métodos y análisis cuantitativo vienen ganando espacio. “Es una buena noticia porque en Brasil, las ciencias sociales siempre tuvieron una gran fragilidad en ese campo, que también se extiende al análisis cualitativo y a estudios con muestras pequeñas”, sostiene Marques, en referencia a iniciativas tales como la Escuela de Verano en Conceptos, Métodos y Técnicas en Ciencia Política y Relaciones Internacionales, ofrecida por la Asociación Internacional de Ciencia Política (Ipsa), el Departamento de Ciencia Política de la FFLCH-USP y el Instituto de Relaciones Internacionales de la USP. También cobran importancia las asignaturas referidas al uso ético de datos. “Es un tema emergente que no sólo apunta prevenir la difusión de datos secretos sobre pacientes o informaciones sensibles a la seguridad pública”, remarca Claudia Bauzer Medeiros. Se corre el riesgo de realizar análisis sesgados porque muchos programas de computación “aprenden”

mientras van procesando los datos. El *software* está desarrollado para identificar patrones a lo largo del tiempo e incorporarlos a su capacidad de análisis. “Ya ha habido casos en que el aprendizaje reprodujo prejuicios en forma inadvertida. En Estados Unidos, se detectó que un programa experimental utilizado por jueces en algunas ciudades para agilizar los fallos era más riguroso con negros y latinos porque tomaba datos de fallos previos como aprendizaje.

El desarrollo de herramientas informáticas que ayudan a analizar grandes volúmenes de datos sobre salud, demografía y violencia promueve estudios sobre procesos sociales que generan aplicaciones en políticas públicas. “El uso de análisis de datos socioeconómicos y demográficos en estrategias de planificación urbana es algo común. La digitalización de datos sobre las oleadas migratorias facilita estudios que ayudan a comprender futuras tendencias inmigratorias”, cita como ejemplo la investigadora del IC-Unicamp.

Un ejemplo de la implicación creciente de las ciencias sociales con el *big data* en Brasil puede verse en el Centro de Estudios de la Metrópolis (CEM), uno de los Centros de Investigación, Innovación y Difusión (Cepid) financiados por la FAPESP. Una de las facetas del centro es la producción y difusión de datos georreferenciados sobre las metrópolis brasileñas. Los organismos públicos producían datos que no quedaban disponibles y finalmente eran apropiados por empresas que cobraban por proveerlos. El CEM adquirió varias bases de datos y digitalizó otras, ofreciéndolas en su sitio web (fflch.usp.br/centrodame-tropole). Al principio, las colecciones no eran lo suficientemente grandes como para encuadrarse en el concepto de *big data*. Esto cambió hace algunos años, cuando el centro desarrolló un banco de datos adecuado para un gran esfuerzo de investigación sobre el estudio de los modelos de desigualdad en los últimos 60 años. Fue necesaria una labor intensa para darle consistencia a los cuestionarios y corregir lagunas en una muestra remanente del Censo de 1960, cuyas tarjetas perforadas se perdieron, y reorganizar la información de los cinco censos posteriores para generar datos comparables. “Eso generó un banco de muchos terabytes de información en un volumen mucho mayor que el tradicional para las ciencias sociales en el país”, dice Eduardo Marques, quien fue director del CEM entre 2004 y 2009. Ese esfuerzo se tradujo en el libro *Trajetórias das desigualdades – Como o Brasil mudou nos últimos 50 anos* (Editorial Unesp, 2015), coordinado por la actual directora del CEM, Marta Arretche, con capítulos escritos por especialistas en temas tales como educación e ingresos, demografía, mercado laboral y participación política. Cada capítulo exigió un procesamiento específico de datos. ■