Reflexiones sobre la construcción automatizada de la información

Responsabilidad mediante algoritmos

James T. Hamilton y Fred Turner

La principal función del periodismo informático (PI) es capacitar a los periodistas para la exploración de grandes cantidades de información, tanto estructurada como desestructurada, en la búsqueda de noticias.
En este texto se extraen las reflexiones de distintos expertos del mundo de la comunicación y la informática con el objetivo de encontrar puntos comunes que proporcionen nuevas herramientas a los periodistas para poder adaptar su profesión, lo más eficientemente posible, al la red.

Palabras clave: periodismo, informática, tecnología, automatización

The main function of computer journalism (CI) is to enable journalists to explore large amounts of information, both structured and unstructured, in search of news.
In this paper draws the reflections of various experts in the world of communication and information technology with the aim of finding common ground to provide new tools for journalists to adapt their profession, as efficiently as possible to the net.

Keywords: journalism, computing, technology, automation

James T. hamilton es profesor de políticas públicas y director del Centro para los Medios de Comunicación y la Democracia, Duke University. (Estados Unidos)
Fred Turner es profesor adjunto y director de estudios de pregrado del Departamento de Comunicación, Stanford University. (Estados Unidos)

En los últimos años, la ubicuidad de la informática ha transformado el paisaje periodístico. Se han socavado los modelos comerciales, reequilibrando el relativo poder de las audiencias y de los periodistas y acelerando la transmisión de información en el mundo entero. Sin embargo, y a la vez que los antiguos modelos de periodismo han ido pasando a un segundo plano, creemos que la informática comienza a ofrecer al periodista una serie de técnicas para facilitar la tarea de informar al público. Las fuentes de datos públicas y privadas se están expandiendo de manera exponencial. Mientras que los defensores de la transparencia hacen presión sobre estas cuestiones, los expertos en informática intentan crear algoritmos con rapidez para que los conjuntos de datos a gran escala tengan sentido. En

la actualidad, las ciencias sociales también trabajan con datos y crean gran cantidad de publicaciones, lo cual significa que se encuentran con los mismos problemas que los periodistas. Creemos que la convergencia de estas áreas de trabajo promete el desarrollo de un nuevo campo: el periodismo informático (PI).

¿Qué es el periodismo informático? En última instancia, son las interacciones producidas entre periodistas, programadores de *software*, informáticos y otros académicos las que deberán contestar a esta pregunta en los próximos años. Por ahora, definiremos el periodismo informático como la combinación de algoritmos, datos y conocimiento de las ciencias sociales, que complementan las funciones de responsabilidad del periodismo. El periodismo informático, de alguna manera, se basa en dos enfoques conocidos, el periodismo de base de datos (PBD o CAR) y el uso de las herramientas de trabajo de las ciencias sociales, defendido por Philip Meyer en su estudio *Precision Journalism: A Reporter's Introduction to Social Science Methods* (Rowman and Littlefield, 2002). Siguiendo estos modelos, el periodismo informático tiene por objeto capacitar a los periodistas para la exploración de grandes cantidades de información, tanto estructurada como desestructurada, en la búsqueda de noticias.

A su vez, el periodismo informático ofrece un nuevo método de presentación de reportajes de seguimiento de los que depende, en gran medida, el buen funcionamiento de la democracia. El término "reportaje de seguimiento-control" abarca tanto la presentación de informes tradicionales de investigación sobre empresas como la cobertura diaria de instituciones clave. Estos tipos de informes tratan de responsabilizar a los dirigentes, desenmascaran malversaciones y sacan a colación las tendencias sociales. Sin ellos, los ciudadanos contarían con muy poca información para la toma de decisiones importantes. Sin embargo, al tiempo que se ha producido un colapso en los modelos tradicionales de periodismo, también han disminuido los incentivos por realizar este tipo de trabajos. El periodismo informático no puede transformar la situación empresarial en la que se encuentra el periodismo contemporáneo, pero sí puede crear nuevos métodos que reduzcan el coste de este tipo de informes, aprovechando mejor el nuevo contexto de la información y, en última instancia, fomentando que el trabajo de seguimiento que realizan los medios salga a flote en el mar de cambio tecnológico en el que nos encontramos.

¿Cómo podría describirse el periodismo informático?

Durante la última semana de julio del 2009, el Center for Advanced Study in the Behavioral Sciences (CASBS) llevó a cabo el curso de verano *Developing the Field of Computational Journalism*. A este seminario, dirigido por James T. Hamilton (Universidad de Duke) y Fred Turner (Universidad de Stanford), asistieron periodistas, investigadores y miembros de ONGs (ver apéndice) para dialogar sobre la evolución de este reciente campo. Este informe describe el punto de vista de los participantes sobre el probable crecimiento y las contribuciones del periodismo informático en el futuro más inmediato.

Los diálogos entre los participantes del CASBS identificaron al menos cuatro áreas de innovación dentro del periodismo informático: técnicas para la transformación de la información y el descubrimiento de patrones en periodismo de investigación; un *digital dashboard* para periodistas; nuevas estructuras sociales y técnicas para la interacción entre los lectores y los periodistas; y avances llevados a cabo en otras disciplinas que pueden apli-

carse al periodismo. Los participantes pusieron de manifiesto que las innovaciones en cada una de estas áreas provienen de un número muy amplio de comunidades y, con frecuencia, pueden involucrar el replanteamiento de los métodos actuales, así como la invención de otros nuevos. En las siguientes secciones damos ejemplos concretos de los métodos que puede que emerjan de estas áreas y los procesos que se seguirían para su desarrollo. Describiremos con detalle las probables innovaciones para transmitir cómo estos logros ayudarían a los periodistas. Asimismo, pretendemos estimular, entre los lectores de este informe, la reflexión acerca de otras formas en las que podrían aprovechar el desarrollo de estas nuevas técnicas.

1. Extracción de información, integración y visualización

Las fuentes de datos públicas y privadas, cada vez más visibles en internet, están aumentando de forma espectacular las oportunidades para presentar informes de control. Para poder beneficiarse de esta situación, los periodistas necesitan una doble asistencia: en primer lugar, deben conocer las maneras de extraer y de integrar la información estructurada de diversas fuentes de datos tales como textos, videos e internet; y, en segundo lugar, es necesario que cuenten con las herramientas con las que explotar y hacer visibles los patrones de los datos extraídos.

1. El desafío de los datos de información

Los periodistas de investigación, a menudo, se encuentran con el desafío de trabajar con registros de datos desarrollados para otros propósitos, con mucha información carente de estructura e incluso repleta de incertidumbres. En un informe de trabajo, la catedrática Sarah Cohen (Universidad de Duke), desarrolló una lista exhaustiva de las herramientas y métodos que ayudarían a los periodistas al descubrimiento de patrones. Los extractos de su estudio, editados a continuación, proporcionan más detalles sobre los desafíos con los que se encuentran muchos periodistas al escribir artículos de investigación

2. Extracción de textos a partir de conjuntos de documentos encontrados de internet

Hay muchos documentos gubernamentales que aún se proporcionan en papel, frecuentemente con amplias secciones omitidas por cuestiones de privacidad o seguridad. Otros son colecciones de documentos encontrados en búsquedas e imágenes de páginas. Los periodistas carecen de métodos para buscar en internet o clasificar en índices estos documentos. Pocos tienen acceso al sofisticado programa de reconocimiento óptico de caracteres (OCR) y menos aún tienen el soporte informático necesario. Una herramienta ideal permitiría a los periodistas, escasos de recursos, introducir documentos en pdf en un servicio de internet que les devuelva una versión con la que puedan hacer búsquedas y que desde Google Desktop, por ejemplo, pudiera añadir a un índice. Este *software* facilitaría la labor de etiquetar documentos con nombres de personas, lugares y fechas distintas.

Ejemplos de áreas en las que esta herramienta habría prestado ayuda periodistas:

El equipo de transición de la administración Obama publicó cartas de muchos grupos de interés que ofrecían consejo al nuevo presidente, en

[1] Un ejemplo lo encontramos en un artículo publicado este verano en la web de la National Transportation Safety Board: [http://www.ntsb.gov]

materias tan variadas como residencias de ancianos o política agrícola. Sin embargo, las cartas no pueden buscarse en internet ni se pueden descargar, lo cual dificultó poder llevar a cabo un análisis de políticas serio cuando comenzó el trabajo de la nueva administración. Algunos organismos cuelgan documentos originales en sus páginas web.[1] Aunque puede usarse la opción "descargar-todo" para introducir los documentos en una carpeta de búsquedas, algunos de estos archivos presentan formatos distintos como imágenes o cartas escaneadas. Los periodistas no pueden distinguir estos formatos, ni saben cómo averiguar si han realizado una búsqueda completa de los documentos.

3. Transcripción de archivos de audio o video

Los comités del congreso, las leyes de los estados federales, los concejos municipales y algunos tribunales proporcionan solo archivos de audio o video. Los periodistas pueden saber aproximadamente lo que se ha dicho, pero tienen que ver o escuchar horas de reuniones para encontrar las partes que quieren utilizar en un artículo o de las que necesitan para entrevistar a una figura política.También utilizan estas audiciones para encontrar nuevas fuentes, pero con frecuencia no tienen buenos índices que indiquen las personas presentes. La transcripción actual de *software* es muy cara y demasiado difícil para que pueda usarse de manera continua. Los investigadores podrían evaluar el *software* que existe hoy para reconocimiento de voz y (discursos de) video, teniendo en cuenta los niveles de exactitud y claridad necesarios en el ámbito periodístico. Entre otras consideraciones, se podría evaluar si el *software* reconoce bien los cambios de la persona que habla, si indica los archivos de audio o video y si puede resumir los términos encontrados. Hay que tener en cuenta que C-SPAN ya ha comenzado a investigar esta tecnología y puede que sea un buen socio adicional de trabajo en este área de investigación.

La falta de transcripciones dificulta la presentación de artículos de las autoridades locales, los tribunales y las autoridades estatales y federales. Las herramientas que hacen falta para reducir el tiempo que se necesita, beneficiarían a los periodistas en casi todas las agencias de noticias.[2]

4. Transcripción de escritura a mano

A menudo, los periodistas de investigación reciben páginas escritas a mano o a máquina respondiendo a peticiones que hacen a registros públicos. El análisis de estos documentos conlleva volver a escribir las repuestas en una base de datos antes de poder analizarlos. Estos documentos son, con frecuencia, informes de contabilidad básica, como la divulgación de información política y financiera o los resultados de inspecciones, por ejemplo, de un restaurante. Los investigadores podrían colaborar evaluando las herramientas para el uso de documentos escritos a mano, en particular, los que se utilizan en la tecnología de extracción de formularios médicos y la gestión de documentos de uso intensivo en los casos legales.

En comparación con otras investigaciones realizadas en diferentes organizaciones de ámbito periodístico, este es uno de los trabajos de recopilación de datos que lleva más tiempo. Pese a la proliferación de los formularios electrónicos, muchos siguen a disposición del público solo en formato de imagen.[3]

[2] Como ejemplo de grabaciones de video que podrían transcribirse e incluirse en búsquedas tenemos WisconsinEye, la versión estatal de C-SPAN, [http://www.wiseye.orgl]

[3] Algunos ejemplos de formularios incluyendo los de divulgación de información financiera, que se pueden encontrar en [http://clerk.house.gov]. Para ver una muestra, visite [http://clerk.house.gov]

[4] Algunos lugares para buscar estas herramientas serían: Google Charts / Google Maps: [http://code.google.com/apischart/] MIT's Simile project: [http://simile-widgets.org/exhibit/] IBM's ManyEyes: [http://manyeyes.alphaworks.ibm.com/]

Encontrar patrones en los datos: visualización de las plantillas

Las visualizaciones de los datos recogidos para la elaboración de artículos llegan, para la mayoría de las veces demasiado tarde, para poder usarlos con eficacia antes de publicarlos. Los investigadores podrían crear varias plantillas básicas de Flash o Web, para visualizar una combinación de tiempo y espacio.[4] Las dos primeras resultan complicadas para la mayoría de los periodistas, y la última tiene demasiadas limitaciones y requiere que se publiquen los datos no confirmados a todo el mundo. Para que las plantillas puedan adaptarse de forma genérica, estas tendrían que permitir muchos tipos de formatos, desde los archivos originales XML o CSV, hasta copiar y pegar una hoja de cálculo, pasando por introducir y editar datos en un formulario. Haría falta que fueran fáciles de actualizar y de personalizar, y tendrían que contar con cierto grado de diseño. Un motivo para usar Flash y Flex es que la mayoría de las organizaciones de ámbito periodístico aún realizan las visualizaciones en este tipo de formatos. Esto permitiría que el trabajo pasara de analizase a publicarse sin problemas, con una sencilla edición.

La destreza para comprender con rapidez un conjunto de datos complicado, sea pequeño o amplio, está ganando importancia en la presentación de informes de funcionamiento de cuentas. Mejorar la coordinación entre las partes de un artículo complejo también es importante a la hora de fomentar la credibilidad en las audiencias.[5]

Herramientas flexibles para establecer cronologías/líneas de tiempo

Cualquier noticia de larga duración precisa una cronología y una línea de tiempo como herramientas de presentación y escritura de informes. Los investigadores, normalmente, utilizan hojas de cálculo Excel o escriben la información de nuevo en un documento de Word en orden cronológico. Sin embargo, no pueden ampliarlo, catalogar eventos para su publicación, añadir o suprimir acontecimientos, ni usarlos con eficacia. La información derivada de un acontecimiento varía de un juicio a una investigación policial, o incluso una reconstrucción de hechos narrados.

Los avances en programación y diseño podrían crear una cronología fácil y estéticamente decente, y una línea de tiempo que permitiera a los periodistas introducir los datos de varias maneras, les diera la posibilidad de editar y exportar la información de diversos modos, les dejara añadir o eliminar personas o tipos de eventos, aumentar y disminuir periodos de tiempo y cargar miles de entradas, si fuera necesario. En la actualidad, Simile's Exhibit y Timeline son las fuentes abiertas disponibles que más se acercan a esta descripción, pero puede que ya haya productos comerciales que se usen de apoyo en litigios o en la aplicación de la ley.

2. El digital dashboard del periodista (Cuadro de Mando digital)

Los periodistas especializados se enfrentan al continuo desafío de buscar información nueva en fuentes de información que les son desconocidas. Al tiempo que el número y los tipos de fuentes de datos digitales aumentan, los periodistas necesitan una herramienta con la que detectar lo que es nuevo e importante de lo que no, en el flujo de información diaria. A la vez, gracias a los esfuerzos de los creadores de *software* corporativo y de la proliferación de las fuentes abiertas, el número y variedad de herramientas con las que contamos para realizar algunas partes de este trabajo está creciendo

[5] Para ver ejemplos actuales, visite: [http://oakland.crimespotting.org/] [http://projects.nytimes.com/crime/homicides/map] [http://www.nytimes.com] [http://www.poynter.org] [http://www.poynter.org]

vertiginosamente. Proponemos la creación de un *digital dashboard* del periodista que pudiera acoger esas herramientas y ayude a los profesionales en su trabajo diario. Los elementos del *digital dashboard* que salieron a colación en la conferencia de CASBS fueron:

1. Noticias de Google a medida para reporteros

Sarah Cohen, en la descripción que hizo de las potenciales herramientas que ayudarían a los periodistas a encontrar nuevas noticias, afirmaba: Los periodistas especializados, el eje central del periodismo de referencia, tienen que estar al día de lo publicado en lo blogs locales, los portales de noticias en internet, los comunicados de prensa que llegan a través del correo electrónico y las páginas web de los gobiernos. Los profesionales intentan supervisar las fuentes, pero muchas repiten los mismos artículos, por lo que necesitan un método rápido para organizar la nueva información encontrada y una manera eficaz para acceder a la fuente original. Los desarrolladores de *software* podrían crear algo como Noticias de Google personalizadas para los reporteros de calle. Su función sería escanear los sitios web seleccionados y hacer un listado de artículos distintos. Solo incluiría en el listado la fuente original y mostraría los sitios web que han creado un enlace a dicha fuente y, así, se determinaría la importancia de esta. Podría percibir los cambios de tiempo, es decir, la última vez que el reportero usó la herramienta.
La construcción de esta herramienta podría incorporar un conjunto de sitios webs y fuentes predeterminadas que constituya un grupo real de fuentes para reporteros locales y periodistas especializados. Mineápolis, San Diego, Chicago o Nueva York son buenos campos de ensayo, ya que albergan muchas organizaciones independientes periodísticas y tienen líneas geográficas imprecisas. El esfuerzo se construiría en torno a un proyecto de fuente de Phase 2 Technology, llamado Tattler y podría comenzar solo con fuentes Rss, antes de albergar más contenido.

2. Seguimiento de las fuentes

Hoy día, los periodistas suelen hacer listas de seguimiento en apuntes con variedad de formatos: hojas de cálculo en Excel, documentos de Word y libretas de direcciones. Una herramienta que hiciera un seguimiento de las fuentes para un periodista especializado, tendría información de contacto grabada en una base de datos. El *tracker* también recogería las noticias de la web, para que cuando una fuente fuera mencionada en un informe de noticias o en un blog, el *tracker* tomara nota y alertara al periodista. El *tracker*, a su vez, podría buscar en los artículos archivados del profesional y así hacer que el contexto y la historia sean visibles al periodista, que será el encargado de considerar sí incluir la contribución de cierta fuente en su artículo, o no.

3. Alertas de información de tendencias y valores extremos

Ahora se puede recoger información diaria actualizada sobre muchos aspectos de la vida de las grandes ciudades. Identificar una noticia basándose en estos datos, sin embargo, puede ser tarea bastante difícil. La creación de límites para aquello que cambia en valores de datos, indicaría si algo anda fuera de lugar o ha variado y, de esta manera, se podría crear una alerta para el periodista cada vez que aparezca una tendencia, una anomalía o cuando una persona en especial o cierta entidad o localidad aparezca en el flujo de datos.

4. Generador de líneas de tiempo

Un generador de una línea de tiempo en el *digital dashboard* del periodista extraería noticias e información de internet y mostraría, al menos, dos cronologías. Para un mismo hecho, habría una línea de tiempo de acontecimientos que trazaría los incidentes específicos mencionados en la cobertura de una noticia específica. Y, una segunda de cobertura, para mostrar el despliegue del tratamiento mediático de una noticia en blogs y otros sitios específicos. Una función de superposición de ambas líneas, permitiría ver cuándo y dónde se presentó la información de ciertos incidentes en la línea de tiempo *Events.*

5. Anotador

El anotador del *digital dashboard* permitiría al periodista ver noticias anteriores, imágenes, referencias e información contextual al elaborar la información. Al redactar la noticia y mencionar ciertas entidades, como un político, por ejemplo, la información contextual aparecería en una ventana lateral cuyo enlace el periodista podría incluir en su artículo. Esta función permitiría que el periodista proporcionara mayor profundización en los contenidos y contextualizara con más facilidad la información. Además, podría dar a conocer la fuente sobre la que se basan las declaraciones del texto del artículo. El Cuadro de Mando Digital facilitaría la tarea de llevar información del ordenador del reportero a la pantalla o diario digital del consumidor.

3. Interacción entre lectores y periodistas

Las herramientas de extracción de texto necesarias para encontrar patrones y el Cuadro de Mando Digital comparten una misma meta: ayudar al periodista a descubrir nuevas áreas para elaborar un buen periodismo. A un nivel más amplio, las tecnologías digitales han cambiado por completo la organización del trabajo de los periodistas y la forma de los artículos que producen. La investigación en periodismo informático debería capacitar a los periodistas para desarrollar nuevos roles sociales y mantener el trabajo de seguimiento de diferentes instituciones. También debería abrir nuevos caminos para contar noticias que, en última instancia, darían a los lectores mejor información y podrían contribuir a que las empresas de comunicación tuvieran más ingresos para mantener su función de control del sistema democrático.

Phil Bennett, ex-director administrativo del *Washington Post* ofreció este ejemplo de cómo el PI podría haber cambiado la presentación del premio Pulitzer del 2007, que ganó la serie de investigación del centro médico Walter Reed Army Medical Center.[6] Una persona con intención de conocer la noticia sobre la desastrosa manera en que los militares atendían al Centro de Salud de Veteranos de Walter Reed, puede acceder a muchos niveles de información. Se podría leer simplemente el texto original, o podrían seguirse los enlaces a los documentos originales, escuchar entrevistas y hacer seguimiento de otras fuentes de información del Centro de Atención a veteranos. Las lecturas revelarían otros intereses del lector y así se le podrían ofrecer otros tipos de información. Esto permitiría que el lector pudiera obtener noticias diferenciadas dependiendo de sus intereses, de modo que se utilizaran las experiencias y tendencias de lecturas anteriores como indicador de los temas que le in-

[6] en: [http://www.washingtonpost.com]

teresan (como las recomendaciones de la librería electrónica Amazon para sus lectores). Al ofrecer a diferentes lectores contenidos ampliamente heterogéneos, se crearía una experiencia más personalizada y, por consiguiente, menos sujeta a la competencia con otros proveedores de noticias.

Bennet señaló que hoy día los proyectos de investigación a largo plazo son como maratones, con la meta en el día de publicación, en el cual se inicia, si el trabajo es destacado, un corto periodo de victoria. Señala, sin embargo, que las publicaciones en internet podrían constituir la mitad del camino. Una serie como la investigación de Walter Reed, por ejemplo, podría convertirse en el centro de atención en la red de lectores interesados en asuntos de veteranos y la atención que reciben. Si el estudio nutre a una comunidad de personas interesadas en la noticia, los lectores podrían utilizar la página web como lugar de debate para la acción que sigue a la investigación. El grupo de personas más interesadas en el tema podría formar una audiencia que fuese el objetivo de anuncios de salud, defensa nacional y asuntos militares.

Los avances en PI podrían, como consecuencia, alterar la manera de transmitir noticias, ya que ofrece, mediante algoritmos, estratos diferentes de noticias que dependen del interés y elección del lector. Esto localizaría el sitio web original, atraería a subcategorías de lectores con amplio interés en un tema, monetizaría su atención con publicidad objetiva y mantendría la captación a la página web. Esta herramienta, por tanto, tiene el potencial de transformar el periodismo de plazos en una fusión de presentación de informes y organización social. Es decir, aprovechando las maneras en que la tecnología digital facilita el diálogo y la reunión de audiencias, el PI podría crear nuevas mezclas de audiencias, reporteros y comentadores. Esto puede que contribuya a que la audiencia pida más un control más férreo de las instituciones y las empresas por parte de los medios ala vez que puede fomentar que los ciudadanos se involucren más en el proceso democrático.

4. Logros con sentido en otras disciplinas

Otra fuente de innovación en el PI son los logros de otras disciplinas, que se pueden transferir para solventar problemas a los que se enfrentan los periodistas de investigación. Los investigadores de campos como la seguridad nacional, las humanidades digitales, la ciencia políticas y la investigación médica se enfrentan a dilemas como la extracción de datos de muchas fuentes de información dispares. Consideraremos cómo podría cada uno de los siguientes proyectos convertirse en la base para las herramientas que necesitan los periodistas:

Ejemplo de investigación de Seguridad nacional:[7]

Rompecabezas: Visualización para el análisis de investigación-. Se trata de un *software* desarrollado por un equipo de investigadores de Georgia Tech que ofrece representaciones visuales de las conexiones entre individuos y entidades que pueden aparecer en muchos conjuntos de documentos. Este *software* ayuda a analizar asociaciones poco probables y a determinar qué documentos sería aconsejable leer si se está interesado/a en determinadas conexiones entre individuos o grupos específicos.

Ejemplo de investigación en humanidades:[8]

El proyecto de Muninn, primera Guerra Mundial, es un proyecto de investigación internacional multidisciplinar cuyo objetivo es traducir la

[7] en: [http://www.cc.gatech.edu/gvu/ii/jigsaw/]
[8] en: [http://www.muninn-project.org]

información de los formularios militares de la primera Guerra Mundial en bases de datos para poder realizar búsquedas. La cuestión que los informáticos deben resolver es cómo usar métodos estadísticos para descifrar respuestas escritas a mano en los formularios usados en este periodo de tiempo. Cualquier avance en este proyecto constituiría una gran ayuda para la profesión periodística a la hora de llevar a cabo análisis en, por ejemplo, formularios de divulgación de información financiera.

Andy Hall, del Centro de Periodismo de investigación de Wisconsin, apunta a que en Wisconsin, por ejemplo, casi 2.700 oficiales al año archivan formularios de divulgación de información financiera que proporcionan datos y cifras sobre inversiones, direcciones y fuentes de ingresos. Los archivos son normalmente copias impresas que con frecuencia contienen respuestas escritas a mano. La Junta de Responsabilidad del Gobierno Nacional no ha tomado medidas para poner los formularios en internet o traspasar la información a una base de datos. Sin embargo, los logros alcanzados en el procesamiento de formularios en el campo de la investigación en humanidades , pueden transferirse con facilidad como herramientas para ayudar al periodista a analizar documentos de divulgación.

Ejemplo de investigación de Ciencias políticas: [9]

Los profesores Frank Baumgartner, Bryan Jones y John Wilkerson han desarrollado métodos para analizar cambios en la agenda de política pública de los EEUU desde la segunda Guerra Mundial. [10] Se trata de un desafío, ya que estudiar las formas en que los diferentes temas emergen en los artículos de legislación y medios de comunicación, requiere clasificar los muchos asuntos mencionados en millones de leyes y artículos. Mientras que codificar a mano una muestra de leyes y artículos fue en su día la manera más factible de proceder para estudiar un cambio de agenda, los logros recientes en la extracción de textos permiten a los investigadores políticos automatizar partes de la tarea de categorizar documentos basados en los temas que debaten. El Proyecto de agendas comparativas usa herramientas de análisis de texto automáticas desarrolladas por el profesor Paul Wolfgang, de la universidad de Temple, para analizar cambios en los programas de emisión en muchos y diferentes países. Este tipo de herramienta podría modificarse para permitir a los periodistas clasificar la forma en que individuos u organizaciones han cambiado su enfoque o prioridades con el tiempo, analizando documentos y declaraciones asociadas con un político o un grupo.

Ejemplo de investigación en Medicina y Ciencias de la Salud: [11]

El ritmo de producción de artículos en campos como la medicina y la biología presenta dificultades para que los investigadores se mantengan al día con las conclusiones principales. La gran cantidad de información médica disponible dificulta la síntesis de los resultados y la detección de resultados inesperados. Cada vez con mayor frecuencia, las herramientas de análisis de contenido se están desarrollando en este área para realizar extracciones de texto que aíslen declaraciones de intenciones, detecten contradicciones y revelen la"información secundaria" que sin ser el enfoque central de un artículo, puede tener importancia. El trabajo de investigación de Catherine Blake, así como su proyecto Descubrimiento basado en la evidencia (Evidence-Based Discovery) de Ciencias de la Salud , financiado por la agencia de gobierno de los EEUU, National Science Fundation (NSF), tra-

[9] en: [http://www.comparativeagendas.org/]
[10] en: [http://www.policyagendas.org/]
[11] en: [http://ils.unc.edu/~cablake/evid/]

ta de utilizar el análisis de contenido para que los investigadores puedan identificar afirmaciones en distintos artículos. Este tipo de herramienta podría ser útil a periodistas que deseen clasificar las declaraciones extraídas de diferentes documentos sobre actividades gubernamentales.

Efectos probables del periodismo informático

En total, los participantes del curso de verano de CASBS identificaron cuatro áreas en las que se podrían aplicar los potenciales logros del PI: técnicas para la transformación de los datos y para encontrar patrones; un Cuadro de Mando Digital para periodistas; un nuevo rol de seguimiento para los lectores y los profesionales; y, por último, la posibilidad de adaptar los logros alcanzados en otras áreas de investigación de vanguardia al ámbito periodístico. Los puntos clave son:

1. Los ordenadores no reemplazarán a las personas

Las nuevas herramientas causarán que emerjan datos e ideas para que los periodistas puedan explorar con más detalle nuevos campos. Son herramientas que se encargan de suplementar, más que de sustituir la labor de los periodistas. En esencia, estas herramientas extraen datos que interesan e involucran al público. Aunque el término "periodismo informático" puede sugerir para algunos la implantación de "periodistas robóticos", solo mediante la interacción de profesionales de la informática y del periodismo, se desarrollarán las técnicas necesarias para que los reporteros encuentren nuevos temas y enfoques para elaborar la información.

2. Las nuevas herramientas fomentarán la participación de nuevos actores en las funciones de control

Las herramientas desarrolladas también cambiarán el ecosistema del periodismo. Al disminuir la dificultad de la actividad de investigación, el periodismo informático contribuirá a extender la función de seguimiento a un conjunto de personas más amplio. Las pequeñas agencias de presentación de informes sin ánimo de lucro, los ciudadanos con un nivel de compromiso alto o las ONGs que participan en debates de política, pueden utilizar estas herramientas de extracción de texto y el *software* del *digital dashboard* (Cuadro de Mando digital) que se presenta en este informe. Mientras que, con frecuencia, la presentación de informes asistida por ordenador se ha considerado como una habilidad exclusiva atribuida a los periodistas de investigación, un objetivo de los algoritmos del PI es que cualquier persona interesada en seguir las actuaciones de instituciones públicas o privadas pueda utilizar estas herramientas sin dificultad.

3. Las nuevas herramientas cambiarán los datos:

Puede que los datos y documentos usados sean imprecisos, contaminados y poco fiables. Como los datos en bruto se comparten con lectores y el proceso de presentación de informes se vuelve más transparente, tendrán que desarrollarse indicadores sobre la calidad de los datos y su procedencia.

4. Las herramientas tendrán que ser una fuente abierta, fácil de usar:

Las herramientas desarrolladas para los periodistas deberán adquirirse de forma libre o tener un coste de adquisición muy bajo, ya que parece poco probable que los periódicos locales y los proveedores de noticias de inter-

net inviertan en la investigación y el desarrollo de estos *software*. Las herramientas tendrán que ser, también, fáciles de utilizar, ya que puede que los periodistas no tengan tiempo o formación suficiente para usar algoritmos complejos.

5. La financiación tendrá que llegar de otras esferas:

La financiación para las innovaciones en PI tendrá que llegar de ámbitos académicos, gobiernos, organizaciones sin ánimo de lucro y fundaciones. Pocas organizaciones del campo de los medios de comunicación están dispuestas a invertir cantidades sustanciales en áreas que no tienen rendimiento económico inmediato.

¿Cómo está evolucionando el campo del periodismo informático? ¿Por qué es importante la respuesta?

En febrero de 2008, el profesor Irfan Essa, de Georgia Tech, convocó la conferencia "Journalism 3G: The Future of Technology in the Field, a symposium on Computation + Journalism". La conferencia[12] reunió a las partes principales involucradas en la creación del nuevo campo de periodistas informáticos: ingenieros informáticos, periodistas e investigadores de la comunicación.
Esta reunión contribuyó a los albores de la cobertura del nuevo campo con artículos como: "Deep Throat Meets Data Mining", [13] "Toxics When Data Are Polluted" [14] y "Can Computer Nerds Save Journalism?". [15]
Profesores como Rich Gordon de Northwestern, Brant Houston de la Universidad de Illinois, e investigadorers de la Facultad de comunicación de Stanford y del Centro para internet y sociedad Harvard's Berkman están buscando proyectos de enseñanza e investigación que puedan contribuir al trabajo conjunto de desarrolladores de *software* y periodistas. En julio de 2009, Sarah Cohen fue nombrada la primera catedrática en periodismo informático de la Universidad de Duke

En febrero de 2010, Georgia Tech organizará una segunda conferencia de informática y periodismo con el objetivo de progresar en la agenda de investigación de este campo. Esta evolución sugiere que puede que los académicos se conviertan en la fuente de innovación que contribuya al avance del desarrollo de las herramientas del PI. Los participantes del curso del CASBS también señalaron la existencia de nuevas organizaciones que forman un "eslabón medio" de proveedores de información al público.

Organizaciones que trabajan con gran variedad de fuentes públicas de datos tales como MAPLight.org, GovTrack.us, OpenSecrets.org, ProPublica.org, EveryBlock.com, entre otras, se han ofrecido con diligencia para mediar entre las bases de datos públicas, los periodistas y el público en general. La mayoría de estas organizaciones son entidades sin ánimo de lucro. Tomadas de una en una representan un conjunto fascinante de experiencias del uso de la informática para informar al público. Unidas, sin embargo, creemos que constituyen una infraestructura emergente como fuentes de información pública. A medida que las instituciones periodísticas se desintegren, los periodistas comienzan a trabajar en unidades más pequeñas o por libre, esta nueva infraestructura representará una fuente informativa clave. Necesitamos, por tanto, comprender con urgencia cómo se está construyendo, cómo afectan las diversas fuentes de financiación a su potencial informativo y democrático y cómo podemos mantener dichos valores democráticos en esta infraestructura. Aún es pronto para mostrar el impacto del periodismo informático, ya que las herramientas para los profesionales y sus colegas,

[12] en: [http://www.policyagendas.org/]
[13] en: [http://www.miller-mccune.com/media/deep-throat-meets-data-mining-875]
[14] en: [http://www.nieman.harvard.edu/reportsitem.aspx?id=100933]
[15] en: [http://www.time.com/time/business/article/0,8599,1902202,00.html]
[16] en: [http://americanpublicmedia.publicradio.org/publicinsightjournalism/]

entre los que se encuentran académicos y periodistas no se han creado. El campo requiere, como eje central, el uso del poder informático para bajar el coste en la elaboración de la información. Hay tres casos recientes que sirven de ejemplo:

1. Visión pública del periodismo: [16]

American Public Media ha colaborado con Minnesota Public Radio y otros para la creación de una Red de visión pública, es decir, una base de datos que contiene información detallada de demografías, intereses y experiencias de más de 70.000 personas. Los reporteros de radio han utilizado la base de datos para asesorar a los oyentes sobre asuntos que conocen y que les interesan. Como APM (American Public Media) dice: "Nuestra red de más de 70.000 fuentes públicas ha contribuido a que encontremos y presentemos artículos sobre la creciente epidemia de obesidad de las áreas rurales, el descenso de los sindicatos y el impacto de la guerra de Irak en familias y soldados". [17] El diario británico *The Guardian,* ante la tarea de examinar los cientos de miles de nuevos informes de gastos presentados por los ministros, inició un *crowdsourcing*. Como si de un juego se tratase, la organización, llevó a más de 20.000 voluntarios a revisar más de 170.000 documentos durante las primeras 80 horas desde que la página web comenzara a funcionar. Los lectores clasificaron los informes de gastos siguiendo las categorías "No interesante", "Interesante pero conocido", "Interesante "y "Para investigar" , los cuales fueron luego investigados por el diario *The Guardian*.

2. Gastos del programa Stimulus a nivel local: [18]

El 30 de octubre de 2009, el gobierno federal puso a disposición de los medios la primera lista de información detallada de los destinatarios de los fondos del programa Stimulus. Los periodistas tendrían que analizar los datos, que se proporcionarán en formatos que no son fáciles de usar ni de asimilar para los reporteros locales.

En un esfuerzo conjunto, que aprovecharían los estudiantes de desarrollo de *software*, los centros de presentación de informes sin fines lucrativos y los editores y periodistas de investigación, Sarah Cohen, (Duke Unversity) , diseñará un mecanismo antes de que la información se divulgue para que cuando la información sea publicada por el gobierno pueda analizarse y distribuirse a los reporteros locales. En una era en la que las organizaciones de medios de comunicación protegen a reporteros y periodistas especializados, el desarrollo del periodismo informático ofrece una nueva forma de expansión para llegar a los profesionales y también a los ciudadanos. La cantidad de datos que se están haciendo públicos a nivel federal va en aumento, lo que implica que las herramientas del periodismo informático tendrán que ofrecer métodos para capitalizar o anticipar la transparencia gubernamental.

Beth Noveck, directora ejecutiva de tecnología de la Agencia de apertura de gobierno de la administración de Obama, describía en sus interacciones del curso la manera en que las páginas web federales, por ejemplo, www.data.gov y los sitios web locales, como www.datasf.org, proporcionarán más cantidad de información para periodistas.

Los *hackers* cívicos como Josh Tauberer, fundador de GovTrack.us, tendrán cantidades cada vez mayores de información gubernamental sin estructurar para organizar y analizar. Jun Yang, que participó en el curso, puso el énfasis en que los nuevos torrentes de información a disposición del público

[17] en: [http://mps-expenses.guardian.co.uk/]
[18] en: [http://www.recovery.gov/?q=content/recipient-reporting]

y la naturaleza del periodismo, que se ve afectada por el paso del tiempo, requieren desafíos que ya son familiares para los ingenieros de la informática. Estos profesionales deben hacer frente a conjuntos de datos inmensos y, con frecuencia, poco fiables, que incluyen la continua consulta y extracción de datos escalables, la procedencia de la información, la optimización del coste-beneficio de la adquisición de la información, la publicación de información que preserve la privacidad y el razonamiento de datos poco fiables.

Rayvon Fouché y Lucy Suchman, que también participaron en los cursos, subrayaron que las herramientas del periodismo informático ejercerán influencia sobre los tipos de organizaciones que terminen desarrollando la función de seguimiento del periodismo. Si los periodistas especializados los van a adoptar, los algoritmos deberán ser fáciles de usar y los datos deberán estar disponibles a gran escala, Además, deberán contar con un diseño que no derroque por completo las prácticas tradicionales de presentación de informes de control.

En el pasado, las técnicas de presentación de informes asistidos por ordenador tendían a formar parte del ámbito de un grupo especializado de reporteros de investigación. Sin embargo, creemos que el periodista ciudadano, los organismos periodísticos sin fines lucrativos y las ONGs que trabajan en cuestiones de responsabilidad gubernamental, también pueden valerse de las herramientas del PI. No obstante, para poder adoptarlas con éxito, Fouché y Suchman señalaron que los futuros periodistas informáticos tendrán que ser innovadores tanto al pensar en la organización del proceso de su trabajo como en las tecnologías que implementan.

Siguientes pasos

El periodismo informático ofrece la posibilidad de bajar el coste de la presentación de informes del funcionamiento de las cuentas, un problema que se agrava en el caso de los medios de comunicación locales, dado su alto coste y la dificultad de monetizarlo. El curso de verano de CASBS subrayó que el desarrollo de las herramientas de extracción de texto y el *digital dashboard* (cuadro de mando digital) del periodista podrían acelerar los hallazgos de los artículos de investigación. La búsqueda de proyectos de configuración que aúnen a personas de múltiples disciplinas, demostrará el grado al que el PI puede llevar el desarrollo de noticias que, de otro modo, no se contarían.

Los participantes del CASBS creían que si se construyeran herramientas efectivas de presentación de informes y de bajo coste, tanto los periodistas profesionales como los periodistas ciudadanos podrían usarlas. Al igual que en muchas cuestiones de los mercados de información, la pregunta principal es quién pagará la creación de estas herramientas. Puede que muchos tipos de actores obtengan un papel en la construcción de este campo:

1. Las fundaciones son un motor en la experimentación de los medios de comunicación. La voluntad de la Fundación Knight porque la nueva Cátedra de Duke fuera ocupada por un experto en periodismo informático, es una clara señal de la fe depositada en este campo. La reciente ronda de ganadores en *Knight News Challenge* incluía proyectos que podrían concluir en el desarrollo de algoritmos e intercambio de datos para los periodistas. El proyecto *Document Cloud*,[19] financiado por la fundación Knight, es pionero en el proceso de transformar amplias colecciones de documentos

[19] en: [http://www.news-challenge.org/winner/2009/document-cloud]

en bases de datos de búsqueda para periodistas. Las fundaciones con particular interés en mantener la presentación de informes de funcionamiento de cuentas y cobertura de asuntos públicos deberían buscar inversiones en proyectos de PI que tuvieran un alto impacto en el rendimiento. Como las herramientas de los informes podrán ser aplicadas con amplitud a muchas áreas geográficas y temáticas, los algoritmos que se desarrollen habrán de ser escalables, tener metas definibles y lograr categorías que puedan medirse en términos de producción de noticias y, en algunos casos, de difusión política.

2. Las agencias gubernamentales, como la Fundación Nacional de Ciencia y la Fundación Nacional para las Humanidades han financiado proyectos de investigación cibernética exitosos en áreas tan diversas como atención médica, terrorismo y literatura. El desarrollo del periodismo informático puede acelerarse, transformando productos que ya han sido desarrollados con el apoyo del gobierno en herramientas que utilicen los periodistas. Puede que las agencias también quieran convocar concursos para la financiación y desarrollo de los *software* que los reporteros y los periodistas ciudadanos podrían usar. Los funcionarios, a su vez, utilizarían las herramientas de extracción de textos y el *digital dashboard*. Las personas que trabajan en agencias de salud pública o política medioambiental podrían usar las herramientas para examinar los temas a los que hacen seguimiento, como parte del interés que les traen los resultados de temas sociales y medioambientales. Los esfuerzos del gobierno por sacar datos en bruto en un formato que puedan utilizar los periodistas, los cuales incluyen el desarrollo de interfaces de programación de aplicaciones (APIs) de las páginas web del gobierno, pueden reducir el coste de la producción de los artículos.

3. Los centros de investigación académica pueden utilizar su poder de convocatoria para reunir a los investigadores de las universidades para que se centren en las herramientas del PI. El Centro para los Medios de Comunicación y la Democracia DeWitt Wallace de Duke, por ejemplo, está reuniendo a periodistas, estudiosos de la comunicación y a informáticos para formar un grupo de presentación de herramientas. Georgia Tech, Harvard, Northwestern, Stanford y la Universidad de Illinois, están llevando a cabo esfuerzos multidisciplinares similares. La mayor parte de estos trabajos no se ha puesto en marcha en las facultades de periodismo. Si las facultades de periodismo utilizaran dotaciones discrecionales para financiar la investigación y el desarrollo en este campo, podrían convertirse en las incubadoras de herramientas que reporteros (y lectores) usarían sin dificultad.

4. Las organizaciones sin fines lucrativos por lo general están contribuyendo a transformar los datos del gobierno en formatos que puedan usarse por la ciudadanía. Cada vez son más las organizaciones sin ánimo de lucro que presentan informes a estos niveles. Este es el caso de MinnPost y del programa de seguimiento que el Centro de Presentación de informes de investigación lleva a cabo en California. Las organizaciones sin ánimo de lucro pueden jugar un papel clave en la evolución del periodismo informático, sobre todo si están dispuestas a compartir los datos que desarrollan con otras agencias de presentación de informes y a asociarse con programas de innovación académica. Los periodistas de investigación

y los editores podrían asociarse con estos grupos que tienen la intención de transmitir sus herramientas de presentación de informes

5. Los desarrolladores de códigos abiertos podrían traducir las herramientas desarrolladas en otros campos a algoritmos periodísticos. Las compañías de programación. a menudo, empujan a sus empleados a dedicar parte de su trabajo al desarrollo de codificaciones de fuentes abiertas. Para muchos desarrolladores de *software*, contribuir a la elaboración de este tipo de *software*, constituye una forma de expresión y de colaborar con otros. El hecho de que la presentación de informes sea posible gracias al periodismo informático, ha comenzado a atraer a desarrolladores que están dispuestos a contribuir al crecimiento de las herramientas de código abierto. La comunidad Drupal, de la universidad de Stanford ya ha comenzado a aunar esfuerzos para crear un *digital dashboard* para periodistas. Nótese que los algoritmos utilizados por los reporteros tendrán que ser fáciles de usar y con un coste reducido, ya que es poco probable que las compañías de los medios de información inviertan en el desarrollo de *software*.

6. Los periodistas y ciudadanos interesados en monitorear el rendimiento institucional determinarán, en última instancia, con cuánta rapidez y extensión se va a desarrollar el campo del periodismo informático. Los mejores programas constituirán, en esencia, las extracciones de textos de interés público. Estos generarán pistas, corazonadas y anomalías para su posible investigación. Este tipo de herramientas permitirán rastrear la información que existe detrás de un patrón de datos, y garantizará que ésta pueda llegar a manos de los periodistas y de otros ciudadanos interesados en hacer un seguimiento acerca del rendimiento del gobierno. Con menos periodistas especializados, sin embargo, quedará a cargo de las fundaciones, agencias gubernamentales, instituciones académicas, organizaciones sin fines lucrativos y desarrolladores de *software* convocar los recursos y la creatividad necesarios para desarrollar las herramientas informáticas que contribuirán a la preservación de la función de vigilancia del periodismo.

El curso de verano del Centro para el estudio avanzado de las Ciencias de la conducta (CASBS) sobre periodismo informático reunió a periodistas, investigadores y representantes de ONGs para centrarse en el desarrollo de este nuevo campo. Los temas de la agenda evocan la amplia naturaleza de las conversaciones que se mantuvieron: el ecosistema del periodista, los datos, el diseño de interfaz de usuarios y de algoritmos, trabajar de forma conjunta con las ciencias sociales y los siguientes pasos

El grupo de trabajo que asistió fue el siguiente:
Phil Bennett, Eugene C. Patterson profesores de periodismo y políticas públicas, Duke University
Catherine Blake, profesora adjunta de la Facultad de postgrado de Biblioteconomía y Ciencias de la Información, Universidad de Illinois en Urbana-Champaign
Sarah Cohen, catedrática de periodismo y políticas públicas, Duke University
Irfan Essa, profesor de la Facultad de Informática interactiva del Instituto de Informática y profesor adjunto de la Facultad de ingeniería informática y electrónica, Instituto de Tecnología de Georgia.
Rayvon Fouché, profesor adjunto de historia, Universidad de Illinois, en Urbana-ChampaignJames

T. Hamilton, Charles S. Sydnor, profesor de políticas públicas y director del Centro para los Medios de Comunicación y la Democracia, y profesor del Centro DeWitt Wallace de Medios de comunicación y Democracia, Duke University
Phil Meyer, profesor emérito de la Facultad de Periodismo y Medios de comunicación, Universidad de Carolina del Norte, en Chapel Hill
Lucy Suchman, profesora del Departamento de Sociología y del Centro de Estudios científicos, Universidad de Lancaster, UK
Joshua Tauberer, desarrollador de *software*, hacker cívico, doctorado en lingüística (en 2010) de la Universidad de Pensilvania
Fred Turner, profesor adjunto y director de estudios de pregrado del Departmento de Comunicación, Stanford University
Jun Yang, profesor adjunto de ciencias informáticas, Duke University

Así mismo, pudimos beneficiarnos de la asistencia de otros académicos y periodistas que efectuaron ponencias para el grupo de trabajo y debatieron sobre la posible evolución en el campo del periodismo informático una vez efectuadas sus contribuciones. Los ponentes fueron:
Jim Bettinger, director de John S. Knight Fellowships, Stanford University
Zach Chandler, especialista en tecnología académica, Stanford University
Louis Freedberg, director de del Center for Investigative Reporting en California y fundador y director de California Media Collaborative
Andrew Haeg, John S. Knight Fellow, Stanford University
Barry Hayes, Google News
Jeff Heer, profesor adjunto de ciencias de la informática, Stanford University
Carl Malamud, técnico, autor, abogado y fundador de Public.Resource.Org
Christopher Manning, profesor adjunto de ciencias de la informática y lingüística. Erudito de Sony, Stanford University
Daniel Newman, co-fundador y director ejecutivo de MAPLight.org
Beth Noveck, jefe adjunto de Tecnología del *Open Government,* Oficina de Política de Ciencia y Tecnología, de la Oficina ejecutiva del presidente de los EEUU