IX EBAM 2017

Curaduría digital y la preservación de contenidos web: creando una colección de tuits sobre la huelga de la Universidad de Puerto Rico

Joel A. Blanco Rivera*

Universidad de Puerto Rico, Recinto de Río Piedras, San Juan, Puerto Rico

Resumen

El 6 de abril de 2017, ocho de los once recintos de la Universidad de Puerto Rico (UPR) comenzaron una huelga en protesta a los recortes al presupuesto de de la Universidad como parte parte de las medidas de austeridad propuestas por el Gobierno de Puerto Rico para enfrentar la crisis fiscal. Con el objetivo de documentar y preservar las voces sobre la huelga reflejadas a través de la web, incluyendo las redes sociales, el autor creó una colección de tuits sobre la huelga de la Universidad de Puerto Rico. Esta iniciativa es parte de un proyecto del curso Introducción a la Curaduría Digital, donde los estudiantes se han enfocado en crear una colección de páginas web sobre la situación de la UPR y la huelga. Este tipo de proyecto ha sido implementado en años recientes por bibliotecas y archivos.

Esta ponencia utiliza el proyecto de captura de tuits sobre la huelga de la Univesidad de Puerto Rico como estudio de caso para examinar el papel de los archivistas en la documentación de eventos políticos y sociales al momento que están ocurriendo. La ponencia coloca las redes sociales como espacios de memoria, y a los archivos como entes de documentación y difusión de memorias. La era digital, y los retos relacionados a la preservación y divulgación de contenidos web, requieren una intervención más activa de parte del archivista para la preservación y difusión de estos espacios de memoria. Por lo tanto, se explora cómo este tipo de proyectos nos llama a re-conceptualizar fundamentos de la archivística, como por ejemplo la valoración, proceso en el cual se establece qué es de valor histórico y por lo tanto amerita su preservación.

Palabras Clave: curaduría digital, archivística, archivos web, redes sociales

* *Dirección electrónica:* joel.blanco@upr.edu

1. Introducción

La curaduría digital se define como todas las actividades involucradas en el manejo de datos nacidos digitales (Abbot, 2008). El objetivo es mantener la autenticidad, confiabilidad, y accesibilidad de los datos por el tiempo que sea necesario (Harvey, 2010, p. 8). Una de sus caracterísitcas principales es que las actividades requieren la colaboración de diferentes actores, incluyendo archivistas, bibliotecarios, científicos en computación e investigadores. El enfoque inicial de la curaduría digital era el manejo adecuado de datos digitales para la investigación científica, área en la cual bibliotecas académicas han desarrollado planes de gestión de datos (*data management plans*) y proyectos de repositorios institucionales (ver Taylor, 2015). Pero la práctica se ha expandido a las humanidades digitales, y en los archivos históricos a la organización, preservación y descripción de fondos documentales cuyos documentos son primordialmente digitales.

Similarmente, la curaduría digital ha sido insertada a las estrategias para la captura y preservación de contenidos web. Los inicios de estas estrategias se pueden trazar a 1996 con la fundación del Internet Archive en California y el proyecto PANDORA de la Biblioteca Nacional de Australia (Alencar-Brayner, 2016, p. 320). Desde ese entonces se han multiplicado los proyectos de archivos web en archivos nacionales, archivos históricos y bibliotecas académicas. A esto se le añade el creciente número de proyectos para la captura y preservación de contenido en las redes sociales. Esta ponencia describe uno de estos proyectos. De enero a julio de 2017, el autor y estudiantes de la Escuela Graduada de Ciencias y Tecnologías de la Información comenzaron un proyecto para la creación de una colección de contenidos web sobre la situación de la Universidad de Puerto Rico. También se implementó un proyecto de captura de tuits sobre la huelga de estudiantes de la UPR, ocurrida de abril a junio.

La ponencia se divide en tres partes. Primero, se discute el tema de la preservación de contenidos web desde la perspectiva archivística, con un enfoque en el valor de estos contenidos como fuentes de documentación histórica y el papel de los archivistas en la selección, preservación y acceso a los mismos. Segundo, se discute el proyecto de las colecciones de contenido web sobre la Universidad de Puerto Rico, enfocándome en la colección de tuits sobre la huelga. Finalmente, se explican los retos principales de este tipo de proyectos, incluyendo consideraciones éticas en la preservación y acceso a contenidos de las redes sociales.

2. Archivística y preservación de contenidos web

Con el aumento vertiginoso de las redes sociales surgen preguntas desde la perspectiva archivística tanto sobre el valor histórico que las mismas tengan como representaciones de las sociedades, y sobre el papel de los archivos históricos en la preservación y divulgación de los contenidos publicados en internet y las redes sociales. En el contexto de la administración pública, archivos nacionales de varios países han desarrollado proyectos para capturar, procesar y dar acceso a páginas web de dependencias gubernamentales. Tal es el caso del Archivo Nacional de Australia, que ha desarrollado una política y procedimientos para que dependencias gubernamentales transfieran sus contenidos web al Archivo Nacional. Similarmente, bibliotecas a través del mundo han desarrollado proyectos para la captura y preservación de páginas web sobre temas particulares. Por ejemplo, en el 2005 la Biblioteca de Catalunya creó el archivo web *Patrimoni Digital de Catalunya (Padicat)*, que incluye páginas sobre acontecimientos de la vida pública catalana (Llueca et al., 2011).

Similarmente, archivistas y bibliotecarios se han insertado en iniciativas comunitarias para la documentación de eventos políticos y sociales, con particular atención en la captura y preservación de contenidos en la web y en las redes sociales. En Estados Unidos, una de estas iniciativas ha sido el proyecto *Documenting Ferguson*, una colección digital que recopila una variedad de documentos relacionados a la muerte del joven Afro-Americano Michael Brown por parte del policía blanco Darren Wilson en Ferguson, Missouri en agosto de 2014. La colección incluye documentos sobre las protestas que ocurrieron luego de la muerte de Brown (ver http://digital.wustl.edu/ferguson/). La iniciativa es una colaborativa, donde personas y organizaciones contribuyen con sus documentos, de manera voluntaria y sin censura, al acervo de la colección (Organization

of American Historians, 2015). Además, durante los días cuando más se intensificaron las protestas, Ed Summers, del Maryland Institute for Technology in the Humanities, creó una colección de tuits sobre los eventos en Ferguson, capturando sobre 13 millones de tuits del período entre el 10 y el 27 de agosto de 2014 (Summers, 2014).

Estos ejemplos ponen de perspectiva la necesidad de una intervención más activa e inmediata de parte de los archivistas para minimizar la pérdida de información a causa de la obsolescencia tecnológica y la inmensa cantidad de información. Sobre lo segundo, según el *Internet Live Stats* a junio de 2017 existían sobre 1.2 billones de páginas web en el mundo (http://www.internetlivestats.com/watch/websites/). Además, se estima que se publican cerca de 6,000 tuits por segundo, lo cual redundaría en 200 billones de tuits por año (http://www.internetlivestats.com/twitter-statistics/#trend). Ian Milligan (2016), Profesor Asistente en el Departamento de Historia de la Universidad de Waterloo (Canadá), explica que en el presente se genera información sobre eventos e ideas a un ritmo nunca antes visto, y que esta información ha sido publicada y compartida por grupos sociales que usualmente han estado ausente en los documentos históricos. Además, estos ejemplos solidifican uno de los roles más importantes de la archivística, la cual es documentar, a través de la gran cantidad y diversidad de documentos, nuestras sociedades. En este contexto, Terry Cook (2014) nos explica que vivimos en un período donde los archivistas "tienen la gran oportunidad de poder documentar la experiencia social y humana con una riqueza y relevancia nunca antes disponible, y con la oportunidad de poder integrar nuestro enfoque en evidencia, memoria, e identidad en un "archivo total" más holístico y vibrante" (p. 113). La web y las redes sociales son parte de esta riqueza de información.

Sin embargo, este caudal de información acentúa los retos relacionados a la selección de documentos considerados de valor histórico (lo que en inglés se conoce como "archival appraisal") y el cual ha sido una función importante en la archivística, particularmente desde mediados de siglo 20. Además, confronta la idea de que los archivos históricos, a través de la custodia de fondos documentales y colecciones, ofrecen la representación de la sociedad y su memoria colectiva. Y es que como explica el archivista sudafricano Verne Harris (2002), los archivos históricos conservan sólo un fragmento de las experiencias sociales (p. 65). Aunque el punto de Harris se presenta en el contexto de la destrucción sistemática de documentos del régimen apartheid en Sudáfrica, su argumento aplica a proyectos de documentación a través de colecciones web y de redes sociales. Siempre existirá, por razones teóricas y prácticas, selección de contenido, lo cual a su vez significa que, conscientemente o no, múltiples voces no formaran parte del archivo. El proyecto sobre la huelga de la Universidad de Puerto Rico, que se discute a continuación, no se exime de esta realidad.

3. Documentación de la Huelga de la Universidad de Puerto Rico

El 5 de abril de 2017, la Asamblea Nacional de Estudiantes de la Universidad de Puerto Rico aprobó un voto de huelga como estrategia para enfrentar recortes al presupuesto de la UPR. Estos recortes son parte de las medidas de austeridad propuestas por el Gobierno de Puerto Rico para enfrentar su crisis fiscal. Al día siguiente, ocho de los once recintos del Sistema de la UPR comenzaron la huelga indefinida, y un noveno recinto se unió a la huelga 6 días después. La huelga se extendió hasta el 8 de junio. Durante este período, redes sociales como Facebook y Twitter se convirtieron en espacios donde no sólo se reportaba sobre los acontecimientos diarios, sino que además personas se expresaban a favor y en contra de la huelga estudiantil. Estos espacios generaron a su vez una variedad de documentos, incluyendo fotos y videos. Además, noticias y opiniones eran publicadas en periódicos nacionales, regionales y universitarios, en blogs, y en páginas web de organizaciones estudiantiles y civiles.

Desde la perspectiva archivística, estos documentos adquieren un valor porque son evidencia de los sucesos relacionados a la huelga, y del contexto político, social y económico del Puerto Rico del siglo 21. Si bien es cierto que documentos públicos del gobierno de Puerto Rico y de la UPR serán fuentes indispensables para la investigación de estos sucesos, la oportunidad de crear colecciones que incorporen las voces de los ciudadanos expande las posibilidades de una representación más amplia de la huelga y de la crisis fiscal en Puerto Rico. Esta fue una de las razones principales por la cual junto a estudiantes del curso Introducción a la Curaduría

Digital decidimos crear una colección de contenidos web sobre la situación de la UPR. Inicialmente el enfoque era en la creación de una colección de páginas web sobre la UPR, pero los diálogos en clase nos llevaron a incorporar estrategias para la selección y captura de contenidos en las redes sociales. De ahí surgió la iniciativa de crear una colección de tuits sobre la huelga en la UPR. Aunque no hemos podido utilizar el software para la captura, organización, y descripción de las páginas web, los estudiantes han realizado tareas de selección de páginas y creación mínima de metadatos que eventualmente serán incorporados a la colección. Para finales de semestre (julio 2017), los estudiantes habrán preparado una guía de creación de metadatos, y una política de desarrollo de la colección. La colección de tuits sí ha sido creada, y es el enfoque del resto de esta ponencia.

3.1. Creación de una colección de tuits sobre la Huelga UPR

Mientras los estudiantes se enfocaron en la planificación para la colección de páginas web, me di a la tarea de crear una colección de tuits sobre la huelga. Para esto utilicé *twarc* (https://github.com/DocNow/twarc), un programado desarrollado por Ed Summers del Maryland Institute for Technology in the Humanities. *Twarc* fue diseñado con el lenguaje de programación Python. A través de una serie de instrucciones, *twarc* realiza búsquedas en Twitter, y genera un fichero en formato JSON con metadatos de todos los tuits recuperados a través de la búsqueda.

Para la creación de la colección fue necesario establecer unos parámetros de búsqueda, por lo que identifiqué dos etiquetas (*hashtags*) utilizadas comúnmente en los tuits sobre la huelga: #HuelgaUPR y #Huelga2017. Con estos parámetros, le doy la instrucción a *twarc* de que realice una búsqueda de tuits que tengan la etiqueta #HuelgaUPR o #Huelga2017:

```
$ twarc search '#HuelgaUPR OR #Huelga2017' > tweets_HuelgaUPR20170606.json
```

El resultado de esa búsqueda es un fichero JSON (**J**ava**S**cript **O**bject **N**otation), un formato de texto que facilita el intercambio de datos a través de computadoras (http://www.json.org/json-es.html). Aunque se pueden leer los contenidos del fichero, el mismo funciona como una base de datos que puede ser convertido en otros formatos para su análisis y visualización.

Es importante indicar que *twarc* solo captura los tuits que son públicos. Además, debido a limitaciones del Twitter Search API, *twarc* recupera tuits que han sido publicados durante un periodo de 7 a 9 días. Por lo tanto, durante los dos meses de la huelga utilicé *twarc* semanalmente. Luego del fin de la huelga, combiné todos los ficheros JSON para crear un solo fichero con todos los tuits. El producto final fue la captura de 30,663 tuits publicados por 8,366 usuarios.

3.2. Análisis y visualización

¿Cómo analizar una colección con sobre 30,000 tuits? Para esto se utiliza otro programado, *twarc-report* (https://github.com/pbinkley/twarc-report), el cual provee herramientas para análisis y visualización de los datos. Por ejemplo, el total de tuits y usuarios se extrajo de una instrucción en *twarc-report* que genera un informe cuantitativo (ver Figura 1). Además de los totales de tuits y usuarios, el informe genera listas de las diez imágenes y los diez enlaces más compartidos entre los tuits almacenados en el fichero. Uno de los beneficios de conocer estos enlaces es que los mismos pueden ser seleccionados para incorporarlos a la colección de páginas web sobre la Universidad de Puerto Rico. El enlace más compartido fue una noticia del 10 de mayo de 2017 sobre una asamblea de estudiantes en el Recinto de Río Piedras donde se decidió continuar la huelga (Torres Montalvo, 2017). El segundo enlace más compartido fue la transmisión por Periscope de una

```
MacBook-Air-4:twarc-report joelblanco-rivera$ ./reportprofile.py projects/tweets_HuelgaUPR20170411-0617-deduplicated.json
Count:                30663
Users:                 8366
User percentiles: ▇▂▁▁▁▁▁▁▁
                  [54, 14, 8, 5, 4, 3, 3, 3, 3, 3]
Has hashtag:          30404 (99.16%)
Hashtags:               625
Hashtags percentiles: ▇▁▁▁▁▁▁▁▁
                  [96, 2, 1, 0, 0, 0, 0, 0, 0, 0]
Has URL:               3585 (11.69%)
URLs:                  1456
URLs percentiles: ▇▂▁▁▁▁▁▁▁
                  [59, 9, 4, 4, 4, 4, 4, 4, 4, 4]
Has Image URL:        10641 (34.70%)
Image URLs:             837
Image URLs percentiles: ▇▂▁▁▁▁▁▁▁
                  [69, 15, 7, 4, 2, 1, 1, 1, 1, 1]
Retweets:             27098 (88.37%)
Geo:                      8 (0.03%)
Earliest:         2017-04-11 01:11:12 UTC
Latest:           2017-06-16 17:54:32 UTC
Duration:         66 days, 16:43:20
```

Figura 1. Parte del informe generado por twarc-report.

manifestación de estudiantes del Recinto de Mayagüez el 18 de abril, donde cerraron el paso de la carretera aledaña al Recinto (https://www.pscp.tv/w/1OyKAojplkLJb#).

También es posible generar gráficas que permiten al investigador identificar los días con mayor actividad en Twitter. Esto a su vez sirve de punto de partida para investigar qué sucedió en esos días. En el caso de los tuits de la huelga, pude generar una gráfica que presenta la cantidad de tuits por día (ver Figura 2). Según la gráfica, el día con mayor cantidad de tuits con etiquetas #HuelgaUPR o #Huelga2017 fue el 25 de abril de 2017 (2,361 tuits). Ese día estudiantes del Recinto de Río Piedras realizaron una manifestación frente al Centro para Puerto Rico, la fundación y archivo histórico de la ex-gobernadora Sila María Calderón, quien se encontraba reunida con el presidente del Senado de Puerto Rico, Thomas Rivera Schatz (Colón Santiago, 2017). Ese mismo día estudiantes del Recinto de Mayagüez llevaron a cabo una asamblea donde ratificaron la continuación de la huelga (Ortega Marrero, 2017).

4. Discusión

Tanto el proyecto de la colección de contenidos web sobre la UPR como el de la colección de tuits sobre la huelga tienen el propósito de capturar y preservar contenidos en el internet sobre el impacto de la crisis fiscal del Gobierno de Puerto Rico en la Universidad de Puerto Rico. En el caso de la colección de páginas web, la misma estará accesible para su consulta. A julio de 2017, los estudiantes de la EGCTI se encuentran en la fase de planeación y selección de páginas. Esperamos trabajar trabajar la fase de catalogación y acceso de septiembre a diciembre de 2017. Por otra parte, la colección de tuits no estará disponible en acceso abierto. Los términos de servicio de Twitter no permite la publicación de las bases de datos, en nuestro caso el fichero JSON. Pero más allá de los términos de Twitter, existen interrogantes sobre tener un balance entre el valor que pueda tener el contenido en Twitter para propósitos de investigación, y la captura de tuits de personas sin tener su consentimiento explícito.

La comunidad archivística que trabaja en estos tipos de proyectos continuan evaluando las implicaciones éticas de tales iniciativas. Si bien es cierto que una herramienta como *twarc* solo captura tuits que son "públicos" eso no necesariamente significa que los mismos se deben copiar y almacenar. Utilizando de escenario la colección de tuits de la UPR, si una persona decide borrar sus tuits o cerrar su cuenta, y esos tuits fueron capturados con *twarc*, los mismos ya no existen en la web pero sí están almacenados en la colección. ¿Tengo derecho a publicarlos? Además, ¿cuáles son las responsabilidades éticas de los investigadores que utilicen estos datos para sus proyectos? Por otra parte, está el punto de vista de que si consideramos la información en la web y las redes sociales como fuentes importantes para estudiar contextos sociales y políticos, como contenidos que

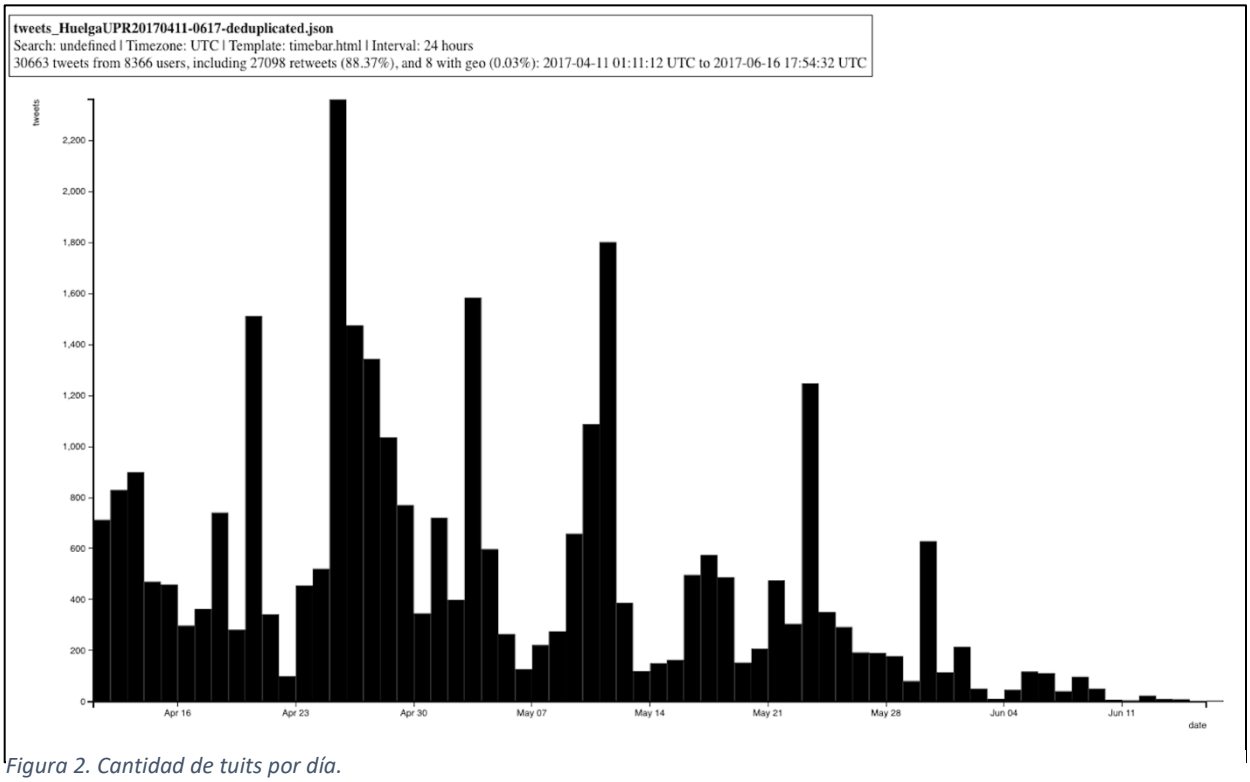

Figura 2. Cantidad de tuits por día.

pueden ampliar, enriquecer, y diversificar la documentación histórica, entonces los archivistas debemos actuar con mayor inmediatez.

Estas consideraciones éticas están siendo estudiadas como parte del proyecto Documenting the Now, el cual tiene como objetivo el desarrollar herramientas que permitan capturar data de las redes sociales (Makiba, 2016, p. 358). *Twarc* es una de las herramientas que forman parte de este proyecto. La comunidad a cargo de esta iniciativa, el cual incluye archivistas, bibliotecarios, historiadores y especialistas en informática, se encuenta analizando maneras de facilitar el consentimiento de los usuarios o alertarlos cuando sus tuits son capturados para una colección (Fields, 2017). Mientras tanto, bibliotecas y archivos evaluan maneras de balancear las consideraciones éticas con la importancia de crear, preservar y diseminar colecciones de contenidos de redes sociales. Por ejemplo, el Cuban Heritage Collection, una división de la biblioteca de la Universidad de Miami, llevó a cabo un proyecto de captura de tuits relacionado al anuncio del Presidente Barack Obama en diciembre 2014 sobre cambios en la política de Estados Unidos hacia Cuba. El acceso a esta colección de tuits es permitido solo a miembros de la Universidad de Miami (http://proust.library.miami.edu/findingaids/?p=collections/findingaid&id=1550). Esta política de acceso es también utilizada por el Bentley Historical Library de la Universidad de Michigan (https://deepblue.lib.umich.edu/handle/2027.42/116594).

5. Conclusión

La primera fase del ciclo de vida de la curaduría digital es la conceptualización, donde se establecen los métodos de captura de datos, las opciones de almacenamiento, y la planificación del proyecto. En el caso de un proyecto de creación de colecciones de contenido web y de redes sociales es fundamental establecer políticas y procedimientos donde se establezcan los lineamientos para la captura, organización, almacenaje, preservación y acceso a las colecciones. Los desarrolladores del programa de código abierto Social Feed Manager

(https://gwu-libraries.github.io/sfm-ui/), compuesto por programadores, archivistas, y bibliotecarios, elaboraron la guía "Building social media archives: collection development guidelines" (https://gwu-libraries.github.io/sfm-ui/resources/guidelines). Esta guía povee una lista de preguntas que se pueden considerar al momento de la conceptualización de un proyecto de este tipo. Las preguntas cubren consideraciones éticas, estrategias de captura de contenido en las redes sociales, y acceso. Este documento, al igual que los proyectos iniciados por bibliotecas y archivos, y nuestro proyecto sobre la UPR, demuestran que la conceptualización y planificación son más importantes, y más complejas, que las decisiones sobre qué tecnologías utilizar para el proyecto.

Finalmente, aún cuando el objetivo de estos tipos de proyectos es ampliar el acervo documental e inclusive diversificar las voces y experiencias reflejadas en los documentos, continuamos documentando un fragmento de la experiencia social. Igual que con otros acervos, la colecciones de contenido web requieren un proceso de selección. Esto pone de perspectiva, tal y como explican Joan M. Schwartz y Terry Cook (2002), que archivistas intervienen e influyen en la conformación de la documentación histórica, y por lo tanto. juegan un papel como mediadores y formadores de memoria histórica.

Referencias

Abbott, D. (2008). What is Digital Curation? DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Recuperado de http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation.

Alencar-Brayner, A. (2016). UK web archive programme a brief history of opportunities and challenges/Programa de arquivo de paginas web no reino unido uma breve historia de oportunidades e desafios/Programa de archivo de paginas web en Reino Unido una breve historia de oportunidades y desafíos. *Revista Digital de Biblioteconomica e Ciencia da Informacao, 14*(2), 318+.

Colón Santiago, N. (2017, 25 de abril). A gritos y golpes la salida de Rivera Schatz de Fundación Sila Calderón. *Noticel*. Recuperado de http://www.noticel.com/noticia/202339/a-gritos-y-golpes-la-salida-de-rivera-schatz-de-fundacion-sila-calderon-video-y-galeria.html

Cook, T. (2013). Evidence, memory, identity, and community: four shifting archival paradigms. *Archival science*, 13, 95-120. doi: 10.1007/s10502-012-9180-7.

Fields, L. (2017, February 24). Why archivists are scrambling to back up the internet. *Deseret News*. Recuperado de http://www.deseretnews.com/article/865673957/The-internet-is-incredibly-fragile-1-heres-how-archivists-are-scrambling-to-save-it.html

Foster, M.J. & Evans M.R. (2016). Libraries creating sustainable services during community crisis: Documenting Ferguson. *Library Management*, 37(6/7), 352-362.

Harris, V. (2002). The archival sliver: power, memory, and archives in South Africa. *Archival Science*, *2*(1), 63-86.

Harvey, R. (2010). *Digital curation: a how-to-do-it manual*. New York: Neal-Schuman Publishers.

Lueca, C., Cócera-Saló, D., Torres, N., Suades-Méndez, G. & De la Vega-Sivera, R. (2011). Al ritmo de tweet: archivando elecciones 2.0. *El profesional de la información*, 20(3), 309-314.

Milligan, I. (2016, December 16). The problem of history in the age of abundance. *The Chronicles of Higher Education*, 63(17), B4+.

Organization of American Historians (2015, August 10). Documenting Ferguson [Blog post]. Recuperado de http://www.processhistory.org/documenting-ferguson/

Ortega Marrero, M. (2017, 25 de abril). Estudiantes votan a favor de continuar la huelga en el RUM. *El Nuevo Día*. Recuperado de https://www.elnuevodia.com/noticias/locales/nota/estudiantesvotanafavordecontinuarlahuelgaenelrum-2314700/

Schwartz, J.M. & Cook, T. (2002). Archives, records, and power: the making of modern memory. *Archival science* 2(1-2): 1-19.

Summers, E. (2014, August 30). A Ferguson Twitter archive [Blog post]. Recuperado de https://inkdroid.org/2014/08/30/a-ferguson-twitter-archive/

Taylor, C. (2015). Research data management: Briefing for library directors. Society of College, National and University Libraries (SCONUL). Recuperado de http://www.sconul.ac.uk/sites/default/files/documents/SCONUL%20RDM%20briefing.pdf

Torres Montalvo, V. (2017, 10 de mayo). UPR Río Piedras decide permanecer en huelga. *Pulso Estudiantil*. Recuperado de http://www.pulsoestudiantil.com/upr-rio-piedras-decide-permanecer-en-huelga/