Revista e-Ciencias de la Información
ISSN: 1659-4142
revista.ebci@ucr.ac.cr
Universidad de Costa Rica
Costa Rica

Blanco-Rivera, Joel Antonio
Tendencias y desafíos en la creación de metadatos en proyectos de archivo de la web
Revista e-Ciencias de la Información, vol. 12, núm. 1, 2022, Enero-Junio, pp. 79-95
Universidad de Costa Rica
San José, Costa Rica

UNIVERSIDAD DE COSTA RICA
EBCI Escuela de Bibliotecología y Ciencias de la Información

e-Ciencias de la Información

Tendencias y desafíos en la creación de metadatos en proyectos de archivo de la web

Joel Antonio Blanco-Rivera

Recibido: 16/03/2021 | Corregido: 11/09/2021 | Aceptado: 16/09/2021

e-Ciencias de la Información, volumen 12, número 1, Ene-Jun 2022
DOI: http://dx.doi.org/10.15517/eci.v12i1.46249
ISSN: 1649-4142

Tendencias y desafíos en la creación de metadatos en proyectos de archivo de la web

Trends and challenges of metadata creation in web archiving projects

Joel Antonio Blanco-Rivera[1]

RESUMEN

Este trabajo tiene como objetivo identificar prácticas y desafíos relacionados al manejo de metadatos en proyectos de archivado web. Se lleva a cabo un análisis de literatura sobre sobre metadatos en archivos de la web publicadas desde 2013, con particular atención a identificar estudios sobre estas prácticas en archivos y bibliotecas. Estos hallazgos apuntan a la necesidad de proveer una mayor contextualización sobre la conformación de un archivo o colección web, así como la importancia de tender puentes entre prácticas bibliotecológicas y archivísticas. Además, para profundizar sobre la importancia de los metadatos en proyectos de archivo de la web, se presenta como estudio de caso la conformación y catalogación de la Colección Web #RickyRenuncia, que forma parte del Proyecto RickyRenuncia, una iniciativa colaborativa y voluntaria para documentar las protestas que llevaron a la renuncia del gobernador de Puerto Rico en julio de 2019. A manera de conclusión, se argumenta sobre la necesidad de conocer más sobre prácticas de archivado web en Latinoamérica y el Caribe, y sobre la posibilidad de establecer redes de colaboración dirigidas a fortalecer la preservación de contenidos web en la región.

Palabras Clave: *archivos web; metadatos; preservación digital*

ABSTRACT

This paper aims to identify practices and challenges related to metadata management in web archiving projects. An analysis of literature on metadata in web archives published since 2013 is carried out, with particular attention to identify studies on these practices in archives and libraries. These findings point to the need to provide further contextualization on the shaping of a web archive or collection, as well as the importance of building bridges between library and archival practices. In addition, to deepen on the importance of metadata in web archiving projects, the conformation and cataloging of the #RickyRenuncia Web Collection, which is part

1 Escuela Nacional de Conservación, Restauración y Museografía "Manuel del Castillo Negrete", Ciudad de México, MÉXICO. Correo: joel_blanco_r@encrym.edu.mx

DOI: https://doi.org/10.15517/eci.v12i1.46249

of the RickyRenuncia Project, a collaborative and voluntary initiative to document the protests that led to the resignation of the governor of Puerto Rico in July 2019, is presented as a case study. By way of conclusion, it argues about the need to learn more about web archiving practices in Latin America and the Caribbean, and about the possibility of establishing collaborative networks aimed at strengthening the preservation of web content in the region.

Keywords: *web archives; metadata; digital preservation*

1. INTRODUCCIÓN

El archivo de la web, entendido como el proceso de selección, acopio y preservación de contenidos web, se ha convertido en una práctica importante dentro de archivos y bibliotecas para lograr la conservación de información digital por su valor administrativo, legal y/o histórico. Ante el exponencial número de proyectos de archivado web ha crecido la literatura teórica y práctica sobre estas prácticas (Brügger y Schroeder, 2017; Summers y Punzalan, 2017; Milligan, 2019; Summers, 2020; Bowyer, 2021). Además, se han llevado a cabos estudios para analizar cómo se llevan a cabo proyectos de archivado web en archivos y bibliotecas, identificando tanto prácticas comunes como desafíos.

Este trabajo tiene como objetivo analizar los procesos de archivo de la web, con particular énfasis en la creación de metadatos para búsqueda y recuperación de información. A través de un análisis de literatura sobre metadatos en archivos de la web se identifican tendencias y desafíos en estas prácticas de preservación digital.

Además, se realiza un estudio de caso sobre la conformación de la Colección Web #RickyRenuncia, con particular atención en la catalogación de los recursos. Esta colección forma parte del Proyecto RickyRenuncia, una iniciativa colaborativa para documentar las protestas en julio de 2019 exigiendo la renuncia del gobernador de Puerto Rico, Ricardo Rosselló Nevares. A manera de conclusión, se proponen potenciales áreas de investigación dentro del archivado web, con un enfoque en Latinoamérica.

2. REFERENTE TEÓRICO

El *International Internet Preservation Consortium* define el archivo de la web como "el proceso de colectar partes de la web, preservando las colecciones en un formato de archivo, y proporcionando acceso a los archivos para su uso" (Web archiving, s.f., párr. 3. Traducción propia). En esta definición se pueden identificar tres características fundamentales del archivo de la web. Primero, la definición especifica que se coleccionan fragmentos de la web, lo cual implica la selección de contenidos a ser preservados. La segunda característica es el almacenamiento en un formato de preservación. WARC (Web ARChive) es el formato de preservación adoptado por la International Organization for Standardization (2017). Y tercero, la definición indica la finalidad del archivo de la web, proveer acceso. Esto a su vez requiere de la creación de metadatos descriptivos para facilitar la búsqueda y recuperación de los recursos. Gilliland (2016) define los metadatos como "la suma total de lo que se puede decir sobre un objeto de información en cualquier nivel

Revista e-Ciencias de la Información
revista.ebci@ucr.ac.cr | http://revistaebci.ucr.ac.cr | ISSN 1659-4142

de agregación" (párr. 3. Traducción propia). Por su parte, Martínez Arellano y Amaya Ramírez (2017) explican que los metadatos descriptivos son un tipo de metadatos que "se utilizan para describir e identificar los principales atributos o características de los recursos de información" (p. 4).

Las prácticas de archivo de la web tienen sus orígenes durante la segunda mitad de la década del 1990. El Internet Archive comenzó a archivar la web en 1996 (Hanna, 2014, p. 82). En marzo de 2003 la UNESCO publicó la *Directrices para la preservación del patrimonio digital*. Como explica Santos Aramburo (2013), este documento, junto a la *Carta para la preservación del patrimonio digital*,

> (...) significó no solo un punto de partida para que distintos países pusieran en marcha estrategias encaminadas a la protección de su patrimonio digital, sino también fue determinante para sensibilizar sobre una situación que se venía produciendo desde hacía años, y cuyas consecuencias negativas aumentaban a medida que no se adoptaban soluciones. (p. 102).

Cuatro meses después de la publicación de las Directrices, julio de 2003, se establece formalmente el IIPC con la participación de las bibliotecas nacionales de once países y el Internet Archive (Toyoda y Kitsuregawa, 2012, p. 1442). Actualmente el IIPC cuenta con representación de 35 países, conformado por bibliotecas nacionales y regionales, bibliotecas universitarias y archivos (Who is the IIPC, s.f., párr. 3).

En términos técnicos, el archivo de la web tradicionalmente consiste en la práctica de web crawling, o rastreador web. Esto significa que un software realiza búsquedas en la web, identifica enlaces y descarga una copia del sitio web. Cuando se realiza en proyectos que tienen un alcance específico, como la creación de una colección temática, las personas encargadas realizan una selección de enlaces que son pasadas al software para que realice el rastreo, captura y almacenamiento. Las personas pueden además especificar al software el nivel de profundidad de la captura (esto es, si desea que se capturen los hiperenlaces que forman parte la página) y la frecuencia. Además, se crean y gestionan metadatos de los recursos acopiados, y se realizan actividades de control de calidad, donde se evalúa la captura realizada por el software.

Esta descripción pone en perspectiva que los procesos de archivado web no son exclusivamente técnicos. El archivado web debe incorporar políticas y procedimientos, además de documentar la toma de decisiones. El Ciclo Vital del Archivado Web elaborado por el Internet Archive muestra claramente estos elementos (Bragg y Hanna, 2013). Como explica Hanna (2014), este modelo "constituye un intento de representar los flujos de trabajo comunes y crear un modelo cuantificable que sirva de referencia para organizaciones para desarrollar o mejorar sus programas de archivado de la web" (p. 83). El modelo ubica a la creación y gestión de metadatos como una actividad fundamental del archivo de la web. Ahora bien, ¿cómo se han trabajado los metadatos en proyectos de archivo de la web? ¿Cuáles has sido los desafíos principales? En las siguientes secciones se atienden estas preguntas y se explica el caso de un proyecto de archivo de la web temático.

3. METODOLOGÍA

El enfoque de esta investigación es el identificar cómo se han llevado a cabo prácticas de creación de metadatos en proyectos de archivo de la web. Es una investigación de corte cualitativo, explicativo, y se divide en dos partes. Primero, se realizó una revisión de literatura sobre archivos de la web para identificar qué se ha publicado sobre metadatos. Se llevaron a cabo búsquedas de fuentes bibliográficas, en español e inglés, en bases de datos Google Académico, Redalyc, así como la consulta de las revistas de archivística Archivaria y Archival Science A partir de esta revisión de literatura, se identificaron estudios realizados sobre prácticas de archivo de la web y se llevó a cabo un análisis documental de estos estudios. El análisis se enfocó en identificar hallazgos sobre organización y descripción de contenidos web. En otras palabras, ¿cómo las instituciones manejan los metadatos de los archivos web? Esto permitió realizar un análisis descriptivo para explicar tendencias en la gestión de metadatos.

Una vez identificadas las tendencias en materia de metadatos, se lleva a cabo un análisis de un estudio de caso, la Colección Web #RickyRenuncia, con el propósito de comparar las prácticas realizadas en este proyecto con las tendencias identificadas en el análisis descriptivo. Una de las características principales del estudio de caso es que permite profundizar sobre uno o varios fenómenos dentro de un contexto particular, por el cual se busca "describir, verificar o generar teoría" (Martínez Carazo, 2006, p. 174). Este análisis del caso de la Colección Web #RickyRenuncia está fundamentado en las experiencias del autor como miembro del equipo de trabajo del Proyecto #RickyRenuncia y responsable de la coordinación del proceso de catalogación de recursos de la colección web.

4. RESULTADOS

4.1 Análisis de estudios sobre archivos de la web

Una revisión de estudios sobre archivos de la web realizadas desde el año 2013 muestra que en términos de creación de metadatos existen variaciones sobre cómo esta se realiza. Entre las razones para esta variación está el que un buen número de herramientas de archivo de la web se enfocan en una actividad específica del proceso, particularmente la captura, lo que requiere la integración de otras herramientas para proveer metadatos descriptivos y acceso a los recursos. Aún con plataformas como Archive-It, que sí incluye la capacidad de proveer metadatos descriptivos y dar acceso, la granularidad varía. Un análisis realizado por el equipo de Archive-It en el año 2013 encontró que un 90 % de las instituciones que utilizan la plataforma generan metadatos al nivel de colección. Sin embargo, solo un 15 % los genera a nivel de ítem (Bragg y Hanna 2013, p. 20). Leisa Gibbons (2016) realizó un estudio de 79 participantes de 13 países (ninguno de Latinoamérica y el Caribe).

Revista e-Ciencias de la Información
revista.eboci@ucr.ac.cr | http://revistaeboci.ucr.ac.cr | ISSN 1659-4142

En relación al tema de metadatos, de un total de 44 participantes que respondieron a esa pregunta, 24 (55 %) indicaron que registran metadatos descriptivos (p. 8). Sin embargo, el estudio no especifica el nivel de descripción. El Web Archiving Environmental Scan de la Biblioteca de la Universidad de Harvard (Truman, 2016), dedica una sección de su informe al tema de búsqueda y recuperación de información, explicando tanto la variedad de estrategias de catalogación como los desafíos sobre este tema que enfrentan los programas de archivo de la web en instituciones. Sobre esto último, bibliotecarios y archivistas entrevistados enfatizaron en la problemática sobre las inconsistencias en los niveles de descripción y en la creación de metadatos. Una de las estrategias para atender este desafío es integrando las descripciones de las colecciones web a los catálogos de bibliotecas y las guías e inventarios de archivos históricos (Truman, 2016, p. 23).

El *Web Archiving Metadata Working Group* de OCLC publicó tres trabajos donde analizan el tema de los metadatos. Estas publicaciones surgieron de dos encuestas, uno dirigido a usuarios y otros a conservadores de archivos web. En ambos estudios se identificó como el desafío principal la necesidad de una estrategia común para la creación de metadatos (Dooley y Bowers, 2018, p. 5). En *Descriptive Metadata for Web Archiving*: *Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*, Dooley y Bowers explican que conservadores de archivos web identifican la necesidad de tender puentes entre las prácticas de descripción bibliotecológicas y archivísticas, una necesidad que "aumenta en importancia por los nuevos tipos de contenido digital que integran nuestras colecciones" (p. 8. Traducción propia). Esto es necesario en gran parte porque los autores encontraron inconsistencias significativas en el uso de estándares para la descripción de archivos web (p. 12). Como propuesta, Dooley y Bowers elaboraron un esquema de metadatos compuesto por 14 elementos. Los elementos son muy similares a Dublin Core. Queda por analizar cómo este esquema es implementado y si logra atender algunas de los desafíos de la descripción de archivos web.

En *Descriptive Metadata for Web Archiving: Literature Review of User Needs*, Venlet, Farrell, Kim, O'Dell y Dooley (2018) apuntan a la necesidad de enriquecer la descripción de los recursos, particularmente en términos de contextualizar la conformación y los contenidos de las colecciones web. En este sentido, las autoras encontraron en esta revisión de literatura que los usuarios desean tener más información sobre la procedencia de los recursos, lo cual permite proveer un mayor contexto sobre su conformación. Esto refleja, explican los autores, "un amplio deseo de transparencia sobre la toma de decisiones relacionadas a la selección de sitios web para su acopio y conformación de colecciones, así como cuán completas son las capturas individuales y los cambios que ocurren a través del tiempo" (p. 4. Traducción propia). Por lo tanto, es necesario que las herramientas utilizadas para los procesos de archivo de la web permitan la integración de información contextual necesaria para el descubrimiento y uso de los contenidos en los archivos web (p.10).

La falta de mayor contextualización sobre las colecciones web es un tema importante discutido por varios autores. Belovari (2017) señala la preocupación de historiadores sobre cómo realizar investigación consultando archivos web, en gran parte por la poca información sobre cómo se seleccionan y organizan los contenidos. Ante esta falta de documentación, expone Belovari, "investigadores no pueden analizar las colecciones en relación a su alcance, significado, y representatividad" (p. 69. Traducción propia). D'Amaro (2019)

enfatiza en la necesidad de que usuarios se acerquen más a los archivos web, "no solo para usarla como fuente para la investigación, sino para que comprendan la creación y la gestión de las colecciones y puedan contribuir, de formas directas e indirectas, a su desarrollo" (p. 270). Este acercamiento, explica D'Amaro, va más allá de comprender el proceso de archivo de la web. Investigadores pueden jugar un papel importante en su conformación, colaborando con conservadores web en la selección de recursos (p. 282).

Para atender esta problemática sobre falta de contextualización, Maemura, Worby, Miligan y Becker (2018) proponen el concepto "procedencia de archivos web" (web archives provenance) como una estrategia de documentación que permita proveer al usuario información sobre cómo una colección web fue conformada. Esta documentación debe abarcar al menos tres elementos: alcance, proceso, y contexto (p. 1230-1231). El alcance se refiere a informar sobre la selección de contenidos, explicando el propósito de la selección, el enfoque de los contenidos, la frecuencia de captura, y qué ha sido excluido. Sobre el proceso, el propósito es documentar la toma de decisiones durante las fases del archivo de la web. Finalmente, el contexto se refiere al marco legal, mandato institucional y políticas que inciden en el archivo de la web.

En resumen, aunque las prácticas de archivo de la web se han llevado a cabo desde finales de los 1990s, continúan los desafíos relacionados a la selección, descripción y acceso a los recursos. Esto se da, por una parte, por la variedad de herramientas de archivo de la web, muchas de las cuales se enfocan en una actividad, particularmente la captura, y por otra parte por la variedad de estrategias de acopio y descripción en instituciones. Ante este panorama nos encontramos además con los constantes cambios en la infraestructura y los contenidos en el internet, con mayor multimedios, y con las redes sociales, que presentan desafíos tecnológicos para su preservación (ver Acker y Kreisberg, 2020). En el caso particular de los metadatos, es importante desarrollar prácticas dirigidas a una mayor descripción de recursos, que a su vez aumente las posibilidades de acceso y uso. Los procesos de archivo de la web abren además las puertas a iniciativas de mayor colaboración, tanto entre archivistas y bibliotecarios, como con otras comunidades interesadas en los archivos web, como lo son investigadores. En el caso de la Colección web #RickyRenuncia, se llevo a cabo esa colaboración entre archivistas y bibliotecarios para la descripción de sus recursos.

4.2 Estudio de caso: Colección Web #RickyRenuncia

El 13 de julio de 2019, el Centro de Periodismo Investigativo (CPI) publicó 889 páginas de un chat que el gobernador de Puerto Rico, Ricardo Rosselló Nevares, mantenía con sus más cercanos colaboradores, entre los cuales se encontraban sus principales asesores y secretarios de algunas de las agencias públicas (Valentin Ortíz y Minet, 2019). El chat incluía mensajes misóginos, homofóbicos, burlas a figuras públicas y organizaciones como la Colectiva Feminista en Acción y a las muertes ocasionadas por el paso del huracán María en septiembre de 2017. Días antes se habían filtrado algunos mensajes del chat, pero esta fue la publicación más extensa, la cual provocó la movilización masiva y protestas en Puerto Rico, Estados Unidos y otras partes del mundo. La noche del 24 de julio de 2019 Rosselló Nevares anunció que renunciaba a su cargo, efectivo el 2 de agosto. Fue el momento cumbre luego de más de dos semanas de protestas en Puerto Rico e internacionalmente, caracterizadas

Revista e-Ciencias de la Información
revista.ebci@ucr.ac.cr | http://revistaebci.ucr.ac.cr | ISSN 1659-4142

por el estribillo "¡Ricky, renuncia!", el cual también se convirtió en la etiqueta principal de las protestas en *Twitter y Facebook*. Rosselló Nevares presentó su renuncia a través de un video transmitido en la página de Facebook de La Fortaleza, como se le conoce a la residencia del gobernador.

Se transmitieron videos del momento en que manifestantes reunidos en la calle Fortaleza del Viejo San Juan, y calles aledañas escucharon el mensaje. Mientras ocurrían las protestas, archivistas en y fuera de Puerto Rico comenzaron a documentar los eventos de manera individual. Esto incluyó el acopio de tuits sobre las protestas por parte del autor y la recopilación de pancartas realizada por la archivista Irmarie Fraticelli Rodríguez durante la marcha masiva del 22 de julio de 2019 (Blanco Rivera, Fraticelli Rodríguez y Ramos, 2020)[2]. Una pregunta de Marisol Ramos a través *Facebook* sobre quiénes estaban documentando las protestas marcó el inicio de las conversaciones entre Ramos, Fraticelli Rodríguez y el autor para crear un proyecto colaborativo, proyecto al que se unió el ingeniero en computación Eduardo Beltrán Feliciano (Fontánez Rodríguez, Ramírez y Gil, 2020). Esta colaboración espontánea dio origen al Proyecto #RickyRenuncia en agosto de 2019, el cual es totalmente voluntario y sin inscripción a una institución en particular.

El objetivo principal del Proyecto #RickyRenuncia es adquirir, preservar y dar acceso a contenidos, primordialmente digitales, relacionados a las protestas exigiendo la renuncia del gobernador de Puerto Rico, cubriendo principalmente el periodo entre el 13 y el 24 de julio. El proyecto tiene tres componentes principales: un dataset de sobre un millón de tuits con las etiquetas #RickyRenuncia y #RickyVeteYa[3], la Colección Web #RickyRenuncia[4], y la planificación de un repositorio digital conformado por materiales digitales donados por personas que deseen compartir sus experiencias y testimonios sobre las protestas. El sitio web del proyecto fue publicado el 1 de septiembre de 2020 (bit.ly/RickyRenunciaProject).

La Colección Web #RickyRenuncia forma parte de *Spontaneous Events Collection*, una iniciativa del Internet Archive donde la organización colabora con grupos o personas interesadas en crear una colección web sobre un evento en la plataforma *Archive-It* (https://archive-it.org/blog/projects/spontaneous-events/). El equipo de trabajo de Archive-It se encarga de crear el espacio de la colección, realizar el acopio de los enlaces seleccionados, e importar los metadatos. Las funciones principales del equipo de trabajo del Proyecto #RickyRenuncia fueron realizar la selección de páginas web y la catalogación de los recursos. La Colección Web #RickyRenuncia cuenta con 418 recursos, disponibles en https://archive-it.org/collections/12491.

2 Las pancartas recopiladas fueron digitalizadas y forman parte del Archivo Digital de Efímera de América Latina y el Caribe de la Universidad de Princeton (https://lae.princeton.edu/?locale=es).

3 La lista de identificadores del dataset están disponibles en el catálogo de tuits de Documenting the Now, https://catalog.docnow.io/datasets/20190930-rickyrenuncia/.

4 En el campo de archivos de la web se han utilizado los conceptos "archivo web" y "colección web" como sinónimos. En el proyecto RickyRenuncia se utiliza el concepto colección porque se alinea mejor a su definición de acuerdo a la archivística. La ISAD(G) define colección como "Conjunto artificial de documentos acumulados sobre la base de alguna característica común sin tener en cuenta su procedencia" (p. 16). Esta característica se da en el Proyecto #RickyRenuncia.

DOI: https://doi.org/10.15517/eci.v12i1.46249

4.2.1 Conformación de la colección web

Es importante presentar algunos apuntes sobre la selección de los recursos de la colección. Así como en las prácticas tradicionales archivísticas y bibliotecológicas se toman decisiones sobre qué conservar, el archivo de la web incluye procesos de valoración y selección de contenidos. En estos procesos se da una intervención tanto de los recursos humanos como los tecnológicos. Como explica Summers (2020), la valoración y selección en proyectos de archivado web "no se pueden llevar a cabo sin la asistencia de tecnologías especializadas y agentes automatizados que recuperan contenidos seleccionados para los archivos, descubren contenidos relacionados, y proveen al archivista un panorama sobre las dimensiones de estas entidades que llamamos páginas web, sitios web, y dominios" (p. 74. Traducción propia). Esto no significa pensar que la automatización de procesos en el archivos de la web es el componente predominante, sino que existe esta interacción humana-tecnológica que permite la conformación, organización y acceso a los contenidos.

En el caso de la Colección Web #RickyRenuncia se vieron estas intersecciones humana-tecnológicas. Es importante destacar que 418 enlaces es un ínfimo número en comparación con el universo de información sobre las protestas que se generó en la web. Por lo tanto debe ser visto como una representación que se vio influenciada por las decisiones tomadas por el grupo de trabajo del Proyecto RickyRenuncia, así como por la capacidad tecnológica de Archive-It para realizar capturas lo mas fidedignas posibles a los enlaces seleccionados. Sobre la intervención de archivistas, Irmarie Fraticelli Rodríguez y Marisol Ramos se enfocaron en realizar búsquedas en internet utilizando las palabras clave Ricardo Rosselló y Manifestaciones Puerto Rico. Para la selección de recursos se dio particular atención a noticias publicadas por medios regionales en Puerto Rico, tanto para preservar contenidos de medios independientes como para documentar las protestas que se llevaron a cabo fuera de la capital, San Juan (Fontánez Rodríguez et al., párr. 11). En relación a la intervención tecnológica, se dieron casos donde Archive-It logró capturar una noticia, pero no el video que la acompaña[5]. Esto último aduce a la realidad de que habrá pérdida de información en los procesos de captura y almacenamiento de contenidos web.

5 Por ejemplo, ver la nota de Noticel "Pedro Rosselló saca la cara por su hijo" (https://wayback.archive-it.org/12491/20190726203437/https://www.noticel.com/ahora/gobierno/pedro-rossello-saca-la-cara-por-su-hijo/1095760383). La noticia incluye un video que no puede ser reproducido en la versión archivada.

Revista e-Ciencias de la Información
revista.eci@ucr.ac.cr | http://revistaeci.ucr.ac.cr | ISSN 1659-4142

4.2.2 Metadatos

El proceso de catalogación de los recursos se llevó a cabo de manera colaborativa. Emanado de una convocatoria lanzada el 30 de julio de 2019 a través de la página de Facebook del Proyecto RickyRenuncia (https://www.facebook.com/rickyrenunciaproject), trece personas se integraron a la primera etapa de catalogación[6]. El grupo incluyó egresados de la Escuela Graduada de Ciencias y Tecnologías de la Información de la Universidad de Puerto Rico (EGCTI) y bibliotecarios y archivistas puertorriqueñas residiendo en Estados Unidos. Cada voluntario tuvo a su cargo la catalogación de 20 recursos, utilizando una hoja de cálculo de Google con los elementos de Dublin Core, utilizado por Archive-It para los metadatos. Los metadatos se incorporaron a una hoja de cálculo principal que fue enviada al grupo de Archive-It para su importación al sistema. Durante la segunda mitad del 2020 realizamos otra fase de catalogación, igual de manera colaborativa. Hasta enero de 2021 hemos logrado catalogar cerca de doscientos recursos.

Evaluando el trabajo de catalogación de la Colección Web #RickyRenuncia en relación a los estudios y hallazgos explicados en la sección anterior, quisiera profundizar en dos aspectos importantes: la contextualización de la colección y el nivel de descripción. Sobre la contextualización, la colección web incluye una descripción general sobre sus contenidos, especificando el periodo que cubre y tipos de recursos que contiene. Comparándolo con el modelo "procedencia de archivos web" de Maemura et al. (2018) solo se puede identificar en la descripción general el elemento de alcance, específicamente sobre el enfoque de los contenidos. Sin embargo, la colección no incluye información sobre criterios de selección y qué ha sido excluido, que también forma parte del elemento alcance. Los elementos proceso y contexto no están representados en la colección web, particularmente en lo relacionado a la toma de decisiones que inciden en la conformación y descripción de la colección web.

En cuanto al nivel de descripción, el objetivo siempre fue proveer la descripción más detallada posible de cada recurso. Con el propósito de mantener uniformidad en la descripción se modificó una guía de metadatos que fue elaborada por estudiantes de la EGCTI en el año 2017 como parte de un proyecto de clase impartido por el autor para crear una colección web sobre la huelga de la Universidad de Puerto Rico (Blanco Rivera, 2017, p. 4). Algunos de los estudiantes que colaboraron en el proyecto de clase luego participaron en la catalogación de la colección web #RickyRenuncia. La guía fue desarrollada utilizando como modelos guías elaboradas en varias instituciones, como por ejemplo la Guía para la creación de metadatos usando Dublin Core de la Universidad de Chile (2009) y el esquema de metadatos de la plataforma de código abierto SobekCM (http://sobekrepository.org/help/metadata), utilizado por la Biblioteca Digital del Caribe. La guía de metadatos de la Colección Web #RickyRenuncia especifica la definición de cada elemento, instrucciones sobre su uso, y ejemplos. El documento facilitó un trabajo uniforme entre el grupo de colaboradores. Además, ayudó a identificar aspectos de la guía que requiriesen revisión, invitando a los participantes a compartir sus observaciones y/o dudas que surgieran durante el proceso de catalogación por medio de correo electrónico.

6 La lista de voluntarios que participaron en la catalogación de recursos está en https://libarchivist.com/rrp/rickyrenuncia/colaboradores?lang=es

De los quince elementos Dublin Core, diez fueron considerados como de uso obligatorio. Para algunos elementos también establecimos reglas particulares para facilitar la claridad y homogeneidad. Dos ejemplos son los elementos "Type" y "Subject". En relación a "Type" elaboramos un vocabulario controlado de cinco tipos de recursos a seleccionar. Por ejemplo, establecimos la distinción entre un editorial y un artículo de opinión, siendo el primero el recurso que expresa opiniones del grupo editorial de un medio de prensa, y el segundo el recurso que expresa la opinión de una persona y es publicado en un medio de prensa. En cuanto al elemento Subject se tomó la decisión de utilizarlo para organizar los contenidos en categorías generales (ej. Gobierno, Movimientos de protesta), en vez de utilizar palabras claves o autoridades como el Library of Congress Subject Headings. La decisión fue tomada por dos razones principales. Primero, por el volumen relativamente bajo de la colección entendimos más viable una organización por categorías, que además reflejan el alcance de la colección en cuanto a los temas principales de sus contenidos. Segundo, se tomaron en consideración las funcionalidades de Archive-It en relación a navegación, búsqueda y recuperación. Aunque la interface tiene un motor de búsqueda, no provee opciones de búsqueda por elemento, pero sí permite navegar por elementos. Por lo tanto, puede ser útil para el usuario escoger de una lista concisa de categorías que agrupa los recursos según el tema principal.

Aunque todavía queda poco más de la mitad de los recursos por catalogar, los avances logrados en este proceso colaborativo de catalogación fueron significativos. Se logró establecer una red de archivistas y bibliotecarios de Puerto Rico y Estados Unidos que facilitó realizar un trabajo fuera de adscripciones institucionales, lo cual ha sido lo realizado por el Proyecto #RickyRenuncia desde sus inicios. Los siguientes pasos deben ir dirigidos a concluir la catalogación y a continuar socializando la colección web para su acceso y uso.

5. CONCLUSIONES

La revisión de literatura y estudios publicados sobre archivado web muestran que aunque el archivo de la web existe desde mediados de los 1990s, todavía falta mucho por recorrer en temas de descripción, acceso y uso. Uno de los desafíos es la necesidad de proveer mayor contexto sobre la conformación de archivos web. El concepto de procedencia de archivos web, con sus tres elementos de alcance, proceso y contexto, provee un buen modelo para contextualizar la conformación y gestión de archivos web. La descripción general de la Colección Web #RickyRenuncia en el Internet Archive incluye información sobre su alcance, pero no se incluye información sobre criterios de selección. En el sitio web del Proyecto RickyRenuncia se especifica que se dió particular énfasis a noticias de medios regionales en Puerto Rico (Web Collection/Colección Web, s.f.).

Por otra parte, se abren posibilidades para tender puentes entre prácticas bibliotecológicas y archivísticas en temas de descripción. Este tipo de colaboración fue identificado como una de las recomendaciones principales en el análisis documental de estudios sobre metadatos en proyectos de archivo de la web. El caso de la Colección Web #RickyRenuncia presenta una manera de lanzar estos puentes de colaboración, a través de un proceso de catalogación colectivo. Esta experiencia ilustra además el valor de la elaboración y uso de guías de metadatos.

Revista e-Ciencias de la Información
revista.ebci@ucr.ac.cr | http://revistaebci.ucr.ac.cr | ISSN 1659-4142

Es de suma importancia realizar un diagnóstico de proyectos de archivo de la web en Latinoamérica y el Caribe. ¿Qué se ha estado trabajando en la región sobre el archivo de la web? ¿Cómo se han desarrollado estos proyectos? ¿Quiénes son los agentes principales que participan? Existen proyectos importantes, como el Archivo de la Web Chilena de la Biblioteca Nacional de Chile donde se pueden identificar lecciones para iniciativas en otros países (ver Aguirre Bello, 2015). Sin embargo, es importante tener un panorama más amplio sobre la realidad latinoamericana en términos de archivo de la web, así como buscar establecer redes de colaboración. El trabajo realizado por la Red Iberoamericana de Preservación Digital de Archivos Sonoros y Audiovisuales (RIPDASA) es un excelente modelo a seguir. Uno de sus componentes principales es el Observatorio de Archivos Sonoros y Audiovisuales de Iberoamérica, cuyo propósito es "dar visibilidad a las instituciones de la memoria que resguardan colecciones sonoras y audiovisuales" (Observatorio de Archivos Sonoros, s.f., párr. 2). Esta visibilidad se da a través un mapa que identifica las instituciones que conservan archivos sonoros y audiovisuales.

Finalmente, es importante reconocer el archivo de la web como un proceso multidisciplinario, donde diversos campos del saber contribuyen con la preservación de contenidos web. El énfasis que da D'Amaro sobre el potencial papel de los investigadores como colaboradores en la conformación de archivos web, acentúa esta importancia. Además, se llevan a cabo procesos de archivo de la web fuera de los confines de los archivos, las bibliotecas, y los museos. En otras palabras, diversos grupos e individuos, se convierten en conservadores web. Desde la historia, por ejemplo, Ian Milligan (2020) argumenta que ante la magnitud de información en la web los historiadores deben realizar trabajos de organización y manejo de contenidos web, desarrollando "habilidades técnicas para organizar datos y encontrar documentos relevantes e importantes dentro del archivo" (p. 240). Este panorama no debe ser visto como un desplazamiento de las funciones de archivistas y bibliotecarios, sino como una oportunidad de reconocer estas prácticas de preservación desde diversos grupos y contribuir con nuestros conocimientos a la preservación de contenidos en la web.

6. AGRADECIMIENTOS

Mi agradecimiento a Marisol Ramos, Irmarie Fraticelli Rodríguez, y Eduardo Beltrán Feliciano, miembros del Proyecto #RickyRenuncia, por el trabajo colaborativo realizado en el desarrollo de este importante proyecto, así como a las y los voluntarios que han hecho posible la catalogación de la colección web.

7. REFERENCIAS

Acker, A. y Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science, 20*(2), 105-123. DOI: 10.1007/s10502-019-09325-9

Aguirre Bello, R. (2015). Archivo de la web chilena: primeros pasos. *Congreso Mundial 2015 de la IFLA*, Ciudad de Cabo, Sudáfrica. Recuperado de http://library.ifla.org/id/eprint/1090/

Belovari, S. (2017). Historians and web archives. *Archivaria, 83*, 59–79. Recuperado de https://archivaria.ca/index.php/archivaria/article/view/13600

Blanco Rivera, J. A. (2017). Curaduría digital y la preservación de contenidos web: creando una colección de tuits sobre la huelga de la Universidad de Puerto Rico. En E*ncuentro Latinoamericano de Bibliotecarios, Archivistas y Museólogos*. Ciudad de México, México. Recuperado de https://www.institutomora.edu.mx/EBAM/2017/Ponencias/Curaduria%20digital%20y%20la%20preservacion%20de%20contenidos%20web%20creando%20una%20coleccion%20de%20tuits%20sobre.pdf

Blanco Rivera, J. A., Fraticelli Rodríguez, I. y Ramos, M. (2020). Documentando lo espontáneo: las protestas #RickyRenuncia. *Archidata: Boletín de la Red de Archivos de Puerto Rico, 18*(1), 13-17. Recuperado de https://archiredpr.files.wordpress.com/2020/11/archidata2020.pdf

Bowyer, S. (2021). The Wayback Machine: notes on a re-enchantment. *Archival Science, 21*(1), 43-57. DOI: 10.1007/s10502-020-09345-w

Bragg, M. y Hanna, K. (2013). *The web archiving life cycle model*. Recuperado de http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

Brügger, N. y Schroeder, R. (2017). *The web as history: using web archives to understand the past and the present*. Londres, Reino Unido: UCL Press

D'Amaro, F. (2019). La memoria digital de España: el archivo web como nueva fuente para la historia del presente. En Moreno Seco, M. (Ed.), *Del siglo XIX al XXI. Tendencias y debates: XIV Congreso de la Asociación de Historia Contemporánea, Universidad de Alicante 20-22 de septiembre de 2018* (pp. 270-285). Alicante, España: Biblioteca Virtual Miguel de Cervantes. Recuperado de http://www.cervantesvirtual.com/obra/del-siglo-xix-al-xxi-tendencias-y-debates-xiv-congreso-de-la-asociacion-de-historia-contemporanea-universidad-de-alicante-20-22-de-septiembre-de-2018-947482/

Dooley, J. M. y Bowers, K. (2018). *Descriptive metadata for web archiving: recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Dublin, Ohio: OCLC Research. DOI: 10.25333/C3005C

Fontánez Rodríguez, C., Ramírez, Y. y Gil, C. (2020). "Somos herederos de todos estos movimientos de protesta que no suceden en un vacío, sino que son toda una historia que hemos estado haciendo desde 1898": entrevista a Marisol Ramos, Joel Blanco Rivera e Irmarie Fraticelli Rodríguez del Proyecto RickyRenuncia (1era parte). Archivoz. Recuperado de https://www.archivozmagazine.org/es/somos-herederos-de-todos-estos-movimientos-de-protesta-que-no-suceden-en-un-vacio-sino-que-son-toda-una-historia-que-hemos-estado-haciendo-desde-1898-entrevista-a-marisol-ramos-jo/

Gibbons, L. (2016). *Web archiving project 2016: preliminary report*. Recuperado de http://leisagibbons.info/wp -content/uploads/2017/03/Webarchivingbriefreport-1.pdf

Gilliland, A. J. (2016). Setting the stage. En Baca, M. (Ed.), *Introduction to Metadata, 3rd ed.* Los Angeles, CA: Getty Publications. Recuperado de http://www.getty.edu/publications/intrometadata/setting-the-stage/

Hanna, K. (2014). El modelo de ciclo de vida del archivado web. En *Anuario AC/E de cultura digital: focus 2014: uso de las nuevas tecnologías en las artes escénicas* [pp- 82-100]. Madrid, España: Acción Cultural Española.

International Organization for Standardization (2017). *Information and documentation — WARC file format* (ISO Standard 28500:2017). Recuperado de https://www.iso.org/standard/68004.html

Maemura, E., Worby, N., Milligan, I. y Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology, 69*(10), 1223–1233. DOI: 10.1002/asi.24048

Martínez Arellano, F. F. y Amaya Ramírez, M. A. (2017). El papel de los metadatos en la Web Semántica. *Biblioteca Universitaria, 20*(1), 3-10. Recuperado de https://www.redalyc.org/articulo.oa?id=28552770002

Martínez Carazo, P. C. (2006). El método de estudio de caso: estrategia metodológica de la investigación científica. *Pensamiento y Gestión,* (20), 165-193. Recuperado de https://www.redalyc.org/articulo.oa?id=64602005

Milligan, I. (2019). *History in the age of abundance? How the web is transforming historical research*. Montreal, Quebec, Canadá: McGill-Queen's University Press

Milligan, I. (2020). La historia en la era de la abundancia: archivos web e investigación histórica. *Historia Y MEMORIA*, (número especial 10 años), 235-269. DOI: 10.19053/20275137.nespecial.2020.11587

Observatorio de Archivos Sonoros y Audiovisuales de Iberoamérica (s.f.). *Red Iberoamericana de Preservación Digital de Archivos Sonoros y Audiovisuales*. Recuperado de https://www.ripdasa.iibi.unam.mx/geoportal/home

Santos Aramburo, A. (2013). El archivo de la web española. *Trama & Texturas*, (22), 101-109. Recuperado de https://www.jstor.org/stable/24391743

Summers, E. y Punzalan, R. (2017). Bots, seeds and people: web archives as infrastructure. En P*roceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 821–834). Portland, Oregon, USA. DOI: 10.1145/2998181.2998345

Summers, E. (2020). Appraisal Talk in Web Archives. *Archivaria, 89(*1), 70–102. Recuperado de https://archivaria.ca/index.php/archivaria/article/view/13733

Toyod, M. y Kitsuregawa, M. (2012). The history of web archiving. *Proceedings of the IEEE, 100*(Special Centennial Issue), 1441–1443. DOI: 10.1109/JPROC.2012.2189920

Truman, G. (2016). *Web archiving environmental scan. Harvard Library report*. Recuperado de http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314

Valentín Ortiz, L. J. y Minet, C. (13 de julio de 2019). Las 889 páginas de Telegram entre Rosselló Nevares y sus allegados. *Centro de Periodismo Investigativo.* Recuperado de https://periodismoinvestigativo.com/2019/07/las-889-paginas-de-telegram-entre-rossello-nevares-y-sus-allegados/

Venlet, J., Farrell, K. S., Kim, T., Allison Jai, O. D. y Dooley, J. M. (2018). *Descriptive metadata for web archiving: literature review of user needs.* Dublin, Ohio, Estados Unidos: OCLC Research. DOI: 10.25333/C33P7Z

Web archiving. (s.f.). *International Internet Preservation Consortium*. Recuperado de https://netpreserve.org/web-archiving/

Web Collection/Colección Web. (s.f.). *RickyRenuncia Project*. Recuperado de https://libarchivist.com/rrp/rickyrenuncia/web-collection-coleccin-web?path=news-and-social-media-archive--archivo-de-noticias-y-contenido-de-redes-sociales&lang=es

Who is the IIPC. (s.f.). *International Internet Preservation Consortium*. Recuperado de https://netpreserve.org/about-us/

e-Ciencias de la Información

2011-2013

Creación de e-Ciencias de la Información como una nueva alternativa, que responde a un contexto marcado por una mayor apertura, flexibilidad y rigurosidad en la publicación científica.

2014-2016

Ingresa a bases de datos de prestigio y calidad como Scielo, DOAJ, Redalyc y otros. Amplía sus horizontes usando como gestor editorial el software OJS y publica en PDF, HTML y EPUB.

HOY

Se encuentra en el cuartil A del UCRIndex y en el Catálogo Latindex con una calificación perfecta, e ingresa al Emerging Source Citation Index de Thomson Reuters.

Revista e-Ciencias de la Información

¿Dónde se encuentra idexada e-Ciencias de la Información?

SciELO · latindex catálogo · Clarivate Analytics · INDIZADA EN redalyc.org · DOAJ

Para más información ingrese a nuestra lista completa de indexadores

¿Desea publicar su trabajo?

Ingrese aquí

O escríbanos a la siguiente dirección

revista.ebci@ucr.ac.cr