Preservación digital de contenidos publicados en la web y las redes sociales

Perla Olivia Rodríguez Reséndiz
Joel Antonio Blanco-Rivera

Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

Este es un resumen legible por humanos (y no un sustituto) de la licencia. Advertencia.

Usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y construir a partir del material

La licenciante no puede revocar estas libertades en tanto usted siga los términos de la licencia

Bajo los siguientes términos:

Atribución — Usted debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.

NoComercial — Usted no puede hacer uso del material con propósitos comerciales.

CompartirIgual — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la la misma licencia del original.

Preservación digital de contenidos publicados en la web y las redes sociales

Colección
Sistemas Bibliotecarios de Información y Sociedad
Instituto de Investigaciones Bibliotecológicas y de la Información

Preservación digital de contenidos publicados en la web y las redes sociales

Perla Olivia Rodríguez Reséndiz
Joel Antonio Blanco-Rivera

Universidad Nacional Autónoma de México
2023

Z701.3 Rodríguez Reséndiz, Perla Olivia
W43R63 Preservación digital de contenidos publicados en la web y las redes sociales / Perla Olivia Rodríguez Reséndiz, Joel Antonio Blanco-Rivera. – México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2023.
xv, 183 p. - (Sistemas bibliotecarios de información y sociedad)

Este libro fue producido en el marco del Proyecto PAPIIT IT 400121 Preservación digital de contenidos publicados en portales web y redes sociales. Del acopio a la difusión de colecciones digitales sobre COVID-19 en México.

ISBN: 978-607-30-8595-3

1. Archivado de la Web. 2. Preservación digital. 3. Redes sociales. I. Blanco-Rivera, Joel Antonio, autor. II. Título. III. ser.

Este libro fue producido en el marco del Proyecto PAPIIT IT 400121 *Preservación digital de contenidos publicados en portales web y redes sociales. Del acopio a la difusión de colecciones digitales sobre COVID-19 en México.*

Diseño de portada: Oscar Arcos
Imágenes: Envato Elements (https://elements.envato.com/es)

Primera edición: diciembre 2023

Instituto de Investigaciones Bibliotecológicas y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-8595-3

Impreso y hecho en México

Índice

Prólogo

¿Qué sería de nuestra sociedad si se perdieran todos los contenidos de la Web? ¿Cómo se escribiría nuestra historia para las futuras generaciones? El interés y la necesidad por salvaguardar contenidos de origen digital en línea nacen casi a la par de la primera publicación en la *World Wide Web* en 1991. Durante casi tres décadas, diversas *comunidades de práctica* se han dedicado a desarrollar un cuerpo consistente de herramientas técnicas-conceptuales para hacer frente a los retos de la preservación de contenidos que se producen, generan, publican y distribuyen a través de la Web.

La preservación digital de la Web y de la información publicada en redes sociales es una tarea ineludible, urgente y compleja. Los contenidos digitales en línea son parte esencial de la huella de la sociedad contemporánea; en ellos se narra, refleja y construye el pensamiento de lo que el sociólogo español Manuel Castells ha denominado como "sociedad del conocimiento"; la Unesco ha reconocido su valor e importancia como parte del patrimonio digital, haciendo hincapié en su inestabilidad inherente y, por tanto, en la apremiante tarea de su salvaguarda.

En su libro *Preservación digital de contenidos publicados en la Web y las redes sociales*, los investigadores Perla Olivia Rodríguez Reséndiz y Joel Blanco-Rivera realizan por vez primera, a partir de un caso práctico en

México y en idioma español, una revisión actualizada y crítica de estas iniciativas; a través de un recorrido que va desde explicar qué es la Web hasta proponer y describir las actividades intelectuales y físicas vinculadas a la preservación, el acceso y el reaprovechamiento de los datos de la web y de las redes sociales, los autores nos ofrecen una guía esencial para el cuidado de esta herencia digital.

En esta publicación se enfatiza la comprensión de contenidos de la Web como parte de nuestro patrimonio digital, con características ontológicas basadas en la transitoriedad, dinamismo, variabilidad e interactividad, cuyo comportamiento a partir de sistemas activos de publicación, flujos de información vinculada e interconectada, así como su funcionalidad determinada por tecnologías y la actividad de múltiples actores sociales, desafía el andamiaje teórico, humano y tecnológico con el que tradicionalmente operan las instituciones de la memoria.

Del universo de información digital en la Web, Rodríguez Reséndiz y Blanco-Rivera abordan las particularidades a las que los archivistas, bibliotecarios y los profesionales de la información se enfrentan para la salvaguarda de los contenidos generados en las redes sociales. Esta publicación invita a reflexionar sobre los principios éticos necesarios que consideren la participación colectiva en la construcción de contenidos, que respeten su privacidad y protejan los datos personales. Problematizan el rol inevitable que tienen las empresas y plataformas privadas en la mediación, acceso y apropiación de nuestras memorias digitales, y describen los retos que enfrentan las instituciones de la memoria para preservar las múltiples dimensiones de los contenidos, apelando a su veracidad, valor y contexto.

Abordando y respondiendo de manera crítica a algunos de los desafíos que supone la preservación de los contenidos de la Web y las redes sociales, este libro denota los principios teóricos, técnicos y éticos que se deben tener en cuenta, así como nuevos roles, habilidades y competencias que esta tarea demanda. Entre los aportes de esta publicación, se destaca la generación de un vocabulario determinado y, sobre todo, la

definición, descripción y alcance de los procesos intelectuales y técnicos asociados con el *Archivado Web*. Rodríguez Reséndiz y Blanco-Rivera sugieren que la implementación de este tipo de proyectos se desarrolle a través de cuatro procesos: 1) Curaduría y selección de contenidos, 2) Acopio, 3) Gestión y almacenamiento digital y 4) Acceso y reaprovechamiento documental.

Esta aproximación integral requiere como primer paso una tarea intelectual en la que se establezcan los criterios, alcances y objetivos para determinar "una política de curaduría y selección", de acuerdo con el contexto, la misión y las obligaciones de la organización. Esta base conceptual determinará qué, cómo, cuándo y cuáles son las herramientas tecnológicas que conformarán el método más adecuado para la formación de colecciones. Los materiales acopiados deberán gestionarse y almacenarse tomando en cuenta los preceptos de la preservación digital sobre autenticidad, fiabilidad y accesibilidad de los objetos, y datos digitales a largo plazo.

Finalmente, los autores abogan a que los documentos de la Web y los datos de las redes sociales, al ser testimonios de la sociedad, deben ser preservados con la misma importancia que tienen otro tipo de documentos, de tal manera que es imperativo establecer criterios éticos, legales, así como estrategias sustentables que permitan, promuevan y aseguren su acceso y uso.

Si bien el *Archivado Web* enfrenta a los archivistas, bibliotecólogos, documentalistas, curadores y profesionales de la información a redefinir sus prácticas, es también una oportunidad para las instituciones que custodian el patrimonio para extender las metodologías e incluso, ampliar sus modelos de uso para el aprovechamiento de tecnologías innovadoras como la Inteligencia Artificial para producir nuevo conocimiento.

Como integrante de una comunidad que se encarga de la conservación del patrimonio artístico digital, agradezco a los autores me hayan permitido incorporar unas palabras en esta obra pionera en muchos aspectos del cuidado de la herencia digital. Celebro el esfuerzo de los

académicos Rodríguez Reséndiz y Blanco-Rivera para generar una guía multidisciplinar, integral y puntual, en la que se articula la polifonía de aproximaciones que a lo largo del tiempo y desde las distintas áreas de conocimiento se han desarrollado para salvar los contenidos de la Web. En manos de Rodríguez Reséndiz y Blanco-Rivera los lectores encontrarán –en este tema relegado de la preservación en nuestro país– un dominio profundo que invitará a más de uno a este mundo apasionante.

Jo Ana Morfín

Introducción

Es sabido que cada época lega su pensamiento y da cuenta de sus acciones como un medio para transmitir información a las generaciones por venir. La información se ha registrado en una amplia gama de soportes producidos con diversos materiales, de acuerdo con el tipo de lenguaje que se plasma: textual, sonoro, audiovisual, imagen fija y multimedios.

Hasta el siglo xx, el libro fue el principal soporte para la conservación y transmisión de conocimiento. Después, con la invención de tecnologías para grabar y reproducir sonidos e imágenes en movimiento se incorporaron otros modos de producir y fijar información, más allá del textual. La información sonora y audiovisual atrajo y cautivó la atención de la sociedad. Primero fue la curiosidad que causó el fonógrafo cuando se presentó como una tecnología capaz de emitir sonidos grabados. Las personas que presenciaron estas audiciones deseaban descubrir quiénes producían los sonidos y cómo se hacía. Más adelante, la radio propició la creación mental de imágenes sonoras a través de la imaginación. Las familias se reunieron alrededor de un aparato de radio para escuchar e imaginar. Por su parte, el cine hizo creer que las imágenes que se veían en la sala se saldrían de la pantalla. Las crónicas de las primeras exhibiciones fílmicas narran cómo ante las imágenes de la llegada de un tren

el público salió corriendo de la sala. Años después, la televisión permitió que se pudieran ver lugares, hechos y personas nunca antes conocidos. Este medio se instaló durante décadas en el imaginario social y solo aquello que se veía existía.

La era mediática, denominada así por la expansión del cine, la radio y la televisión como medios de masas, conllevó a la producción de documentos sonoros (programas de radio, música, grabaciones orales, etcétera) y audiovisuales (filmes, programas de televisión, videorregistros, entre otros). Así, la historia del siglo pasado fue narrada y documentada también con sonidos e imágenes.

El reconocimiento social del valor informativo, documental y patrimonial de los sonidos e imágenes fijas y en movimiento fue tardío. La relevancia de este tipo de materiales como recurso de información, patrimonio y documento fue avizorada apenas a finales del siglo pasado y las tareas para la salvaguarda de este tipo de materiales son recientes. Es posible que, por ello, se hayan borrado y tirado a la basura miles de documentos sonoros y audiovisuales.[1]

Aún sin resolver los complejos desafíos de la preservación de las colecciones sonoras, audiovisuales y fotográficas, la tecnología digital potenció la creación de información digital a través de nuevos tipos documentales: la World Wide Web, también nombrada la Web y las redes sociales.

En estricto sentido, las redes sociales se insertan en la Web como plataformas y aplicaciones mediante las cuales se pueden forjar conexiones personales entre los usuarios para compartir información a través de mensajes personales, publicación de audios, fotografías, videos, formar grupos de interés, entre otros.[2] Cuando nos referimos a las redes sociales tenemos en mente plataformas como Facebook, Twitter, LinkedIn o aplicaciones como x (antes Twitter), Instagram o de forma reciente Threads.

---

1 Perla Olivia Rodríguez Reséndiz, *Estado de la preservación digital en los archivos sonoros y audiovisuales de Iberoamérica* (Quito: Universidad Andina Simón Bolívar, Sede Ecuador, 2020), 96.

2 Marcos Ros-Martín, "Evolución de los servicios de redes sociales en internet", *El profesional de la información* 18, n.° 5 (2009): 552-557. DOI: 10.3145/epi.2009.sep.10.

Desde hace tres décadas la historia y el acontecer del mundo se narra día a día a través de Internet. Quien en un siglo desee conocer las motivaciones, creaciones y desatinos del siglo XXI deberá recurrir a la información digital publicada en forma de páginas web y distribuida en redes sociales.

La Web podría ser estudiada en un futuro como la más grande biblioteca virtual de la humanidad, cuya naturaleza es efímera. Por otra parte, la información que se difunde a través de redes sociales cada día da cuenta de las ideas, creencias, verdades a medias y mentiras que definen a la posverdad como fenómeno social. Y este contenido es al mismo tiempo un invaluable recurso de información con vastas posibilidades de uso en la investigación científica y la docencia.

Son contadas y poco conocidas las tecnologías para el acopio, la gestión y el acceso al material documental y conjunto de datos que se publican en la Web y en las redes sociales. Además, la capacitación de profesionales de la información es incipiente y los programas de estudio a través de los cuales se les forma para la preservación de este tipo de materiales son escasos o bien están desactualizados. En consecuencia, la preservación digital del material publicado en la Web y las redes sociales deviene innecesaria e incomprensiblemente frente al tratamiento documental de materiales que por tradición se salvaguardan en las bibliotecas, archivos y museos.

Por lo tanto, la salvaguarda de este tipo de materiales se percibe como incierta. Da la impresión de que se trata de una tarea infinita que nunca concluirá y por ello, la motivación para emprender la preservación de la Web está llena de dudas. Algunas de las preguntas que se formulan en torno a este tipo de materiales son: ¿Por qué se deben preservar las páginas web y los materiales publicados a través de las redes sociales? ¿En qué institución de la memoria debe recaer la responsabilidad de preservar este tipo de materiales? ¿Debe ser una tarea que se inserte en el trabajo cotidiano de archivos y bibliotecas? ¿Cómo se debe llevar a cabo la salvaguarda de este tipo de materiales? ¿Cuáles son los procesos que deben ponerse en marcha para preservar este tipo de contenidos?

Las respuestas a las preguntas anteriores son complejas e involucran diversas perspectivas. Esta publicación pretende motivar a sus lectores para que al concluir puedan tener respuestas a estas interrogantes. Y al mismo tiempo, se pretende que los lectores formulen otras preguntas en torno al Archivado Web y de redes sociales.

Este libro fue escrito como resultado del proyecto *PAPIIT IT 400121 Preservación digital de contenidos publicados en portales web y redes sociales. Del acopio a la difusión de colecciones digitales sobre COVID-19 en México.* Participaron en esta iniciativa seis investigadores, quienes con una perspectiva resiliente durante la pandemia que asoló a la humanidad en el siglo XXI, pusieron en marcha esta iniciativa para usar herramientas de acopio y preservación de información publicada en portales web y en la red social Twitter sobre COVID-19 en México.

El grupo de investigación fue integrado por los doctores Jenny Teresita Guerra, Jonathan Hernández Pérez, Perla Olivia Rodríguez Reséndiz (coordinadora del proyecto), del Instituto de Investigaciones Bibliotecológicas y de la Información (IIBI), de la Universidad Nacional Autónoma de México (UNAM), y por el doctor Joel Antonio Blanco-Rivera, de la Escuela Nacional de Conservación, Restauración y Museografía "Manuel del Castillo Negrete" (ENCRYM), perteneciente al Instituto Nacional de Antropología e Historia (INAH), y se contó con el apoyo técnico académico de la maestra Dafne Citlalli Abad Martínez, del IIBI, y de la doctora Carolina Silva Bretón, del Instituto de Investigaciones Bibliográficas (IIB), UNAM.

Este libro es resultado de un esfuerzo de colaboración y cooperación internacional. Lo que los lectores encontrarán en las siguientes páginas es fruto tanto de disertaciones teóricas y conceptuales, como de la praxis del uso de herramientas tecnológicas para el Archivado Web y de redes sociales, empleadas durante más de dos años. Esta publicación estaría incompleta sin agradecer el invaluable apoyo que proporcionaron Lorena Ramírez e Ilya Kreymer de Web Recorder, quienes contribuyeron mediante la asesoría para llevar a cabo prácticas de acopio y recopi-

lación de la Web. Además, resulta necesario agradecer el invaluable apoyo de Diego Pino de la Biblioteca Metropolitana de Nueva York, gracias a quien fue posible poner en marcha un archivo digital con los materiales acopiados durante la pandemia. En este tenor deseamos agradecer el apoyo de Jefferson Bailey de Internet Archive por compartir su visión y perspectivas en relación con el Archivado Web y de redes sociales.

Durante la etapa final de la escritura de este libro, en julio de 2023, Twitter cambió su nombre a X, diez meses después de ser adquirido por Elon Musk. En el capítulo II describimos este cambio, ubicándolo en el contexto histórico de la evolución de las redes sociales. En el resto del volumen utilizamos Twitter, debido a que fue la plataforma a través de la cual se realizó el acopio de contenidos sobre la pandemia en México, proyecto que sustentó el desarrollo de esta publicación. Los cambios que ha tenido la plataforma, en términos de políticas de acceso y acopio de datos, han provocado reconsideraciones sobre prácticas de archivado de redes sociales que se abordan en el libro.

Conviene señalar que la presente obra está dirigida a bibliotecarios, archivistas, documentalistas, profesionales de la información e investigadores, con el propósito de poner a su disposición los procesos que intervienen en la preservación digital de los recursos de información publicados en la Web y las redes sociales, toda vez que poseen el carácter de documentación histórica.

La elección de herramientas de acopio, conservación y acceso a contenidos digitales proporcionó experiencia y conocimiento que se comparten para las bibliotecas y archivos que tienen ante sí el desafío de preservar este tipo particular de recursos de información sujetos a la denominada "vulnerabilidad digital".

Los autores

I. El archivado web

1.1. ¿Qué es la Web?

La Web es un artefacto cultural del ecosistema digital del siglo xxi. Opera como un sistema dinámico a través del cual se publica y proporciona acceso a información hipermedia por medio de Internet[3] en un espacio virtual poblado de contenidos digitales multimedia en diferentes lenguajes, cuyo acceso multiplataforma, expansión y crecimiento es continuo.[4]

Puede ser observada como un medio y a la vez como un contenedor de información. Su naturaleza es dinámica y está determinada por las áreas virtuales en las cuales se publican y distribuyen contenidos en diferentes lenguajes (textual, sonoro, audiovisual y fotográfico) que en conjunto conforman el lenguaje hipermedia.

Los contenidos se articulan y relacionan a través de vínculos o *links*. Se determina así la estructura narrativa hipermedia de la Web. La actualización y dinamismo de muchas páginas web y la interconexión de los contenidos son cualidades innovadoras que no tienen otros tipos de materiales informativos, además, ofrece experiencias inmersivas y de complementariedad.

3 Julien Masanès, "Web Archiving: Issues and Methods", en *Web Archiving*, editado por Julien Masanès (Nueva York: Springer, 2006).

4 Javier Guallar y Javier Leiva-Aguilera, *El content curator. Guía básica para el nuevo profesional de Internet* (Barcelona: Universitat Oberta de Catalunya, 2013), colección El profesional de la información 24.

Como sistema activo de publicación se sustenta en la combinación de tres estándares: URI, HTTP y HTML y SGML DTD.[5] El Uniform Resource Identifier (URI) es el estándar a través del cual se identifican los recursos en la Red de forma unívoca en Internet.[6] Dicho de otra forma, es el nombre del espacio que ocupa un objeto digital en Internet. La diferencia del URI con el Uniform Resource Locator (URL) es que este último hace referencia a recursos que pueden variar en el tiempo. El Hyper Text Transfer Protocol (HTTP) es el protocolo de comunicación e interacción entre el cliente o usuario y el servidor para la transferencia de información a través de archivos (XML, HTML…) en la Web. El HTML y la Standard Generalized Markup Language (SGML) son lenguajes a través de los cuales se representa y mantiene la estructura lógica de las partes de un documento hipertextual. Es el lenguaje empleado para marcar y definir la presentación de las páginas en los navegadores.

La Web existe en Internet, un sistema de información digital en constante evolución. Esta característica determina una dependencia con el creador del contenido quien puede borrar, agregar o modificar la información publicada en una página web.

La naturaleza efímera de la Web conlleva a considerar que en cualquier momento la información publicada puede desaparecer o ser alterada. En consecuencia, su preservación debe tomar en cuenta esta naturaleza cambiante y por ello, el Archivado Web es el proceso a través del cual se separa el contenido del sistema de información original y se garantiza que la información pueda ser resiliente a los cambios y evolución de la Web.[7]

Así a través del Archivado Web y la preservación digital se asegura la permanencia de la información digital, de tal manera que pueda ser leída, escuchada o vista, de acuerdo con las características lógicas, estructurales y tecnológicas de los contenidos con las que fueron originalmente

5 Masanès, "Web Archiving".

6 Tim Berners-Lee *et al*., "The World-Wide Web", *Communications of the ACM* 37, n.° 8 (1994), https://composingdigitalmedia.org/digitaliteracy/docs/p76-berners-lee.pdf.

7 Masanès, "Web Archiving".

creados; sin depender de la obsolescencia de tecnologías, protocolos, formatos digitales y aplicaciones con las que fue producida.

El Archivado Web es disruptivo. A diferencia de otros procesos documentales que se abocan al tratamiento de materiales en un lenguaje determinado, este forja documentos de origen digital que comunican información en diversos lenguajes (textual, sonoro, audiovisual y multimedia) y vínculos entre diferentes tipos de contenidos. A través de las técnicas de Archivado Web, una página web deviene en un documento de origen digital que puede ser preservado. Por lo tanto, en su preservación son insuficientes las tradicionales nociones de acopio, organización, conservación y acceso a contenidos empleados por las instituciones de la memoria.

1.2. La Web: omnipresente y efímera

La publicación de información digital en la Web es imparable. Desde el 6 de agosto de 1991 hasta abril de 2023 se contabilizan 1 500 millones de sitios web en Internet.[8] En estos datos no se incluye a la Web profunda, cuyo volumen se estima que es entre 400 y 500 veces superior al de la Web de la superficie.[9]

La Web fue calificada hace más de una década, cuando aún estaba en proceso de expansión, como el documento más grande jamás escrito, cuyo volumen de información hasta entonces era 50 veces mayor a la cantidad de textos conservados por la Biblioteca del Congreso de los Estados Unidos.[10]

La Web es el cristal a través del cual se puede observar el devenir de la humanidad y donde se publica información de todo tipo de temáticas, en todos los idiomas y por parte de prácticamente todas las personas. Se expresan desde informaciones periodísticas, aportaciones científicas,

8 "Total number of websites", Internet Live Stats, consultado el 28 de junio de 2023, https://www.internetlivestats.com/total-number-of-websites/.

9 Peter Lyman y Hal Varian, "How Much Information? 2000", Universidad de California, publicado el 18 de octubre de 2000, http://www.sims.berkeley.edu/research/projects/how-much-info/.

10 *Idem.*

manifestaciones culturales, posicionamientos políticos, propuestas de entretenimiento hasta perspectivas ideológicas y religiosas.

Tienen cabida quienes desean difundir servicios, vender productos, dar a conocer la estructura organizacional de sus instituciones, promocionar acciones de gobierno, entre otros. Incluso en la actualidad, que ha sido caracterizada como la era de la posverdad, es una herramienta a través de la cual se difunden *fake news* y verdades a medias. Estos usos, entre otros, denotan la apropiación de Internet por parte de la sociedad, en contraste con el origen militar que tuvo esta tecnología a través de ARPANET, a finales de los años 60.[11]

Como objeto de estudio atrae la atención de científicos de diferentes disciplinas científicas, sociales y humanísticas. Para disciplinas como la archivología, la bibliotecología, la comunicación y los estudios de la información interesa estudiar a la Web como un medio y como un contenedor de información digital porque advierte, entre otras, la irrupción de nuevas modalidades de información, tipologías de recursos de información, documentos y formas de patrimonio digital; y porque su permanencia es incierta.

La publicación de páginas web se aceleró a partir de 2017.[12] Prácticamente todos los sectores de la sociedad publican y consultan información, aunque no necesariamente la actualizan. Por lo cual, el 75 % de estas publicaciones está inactivo[13] y muchas páginas desaparecen.

Al respecto, Peter Lyman señaló en el año 2000 que el 44 % de los sitios publicados en 1998 ya no se encontraron en 1999.[14] Por su parte, en un estudio publicado en 2017 por Miguel Costa, Daniel Gomes y Mário J. Silva se advirtió que el 80 % de las páginas publicadas no están disponibles en su versión original en un año; el 13 % de las referencias de artículos académicos publicados en la Web desaparecen en 27 meses

11 Adrian Brown, *Archiving Websites: A Practical Guide for Information Management Professionals* (Reino Unido: Facet Publishing, 2006).

12 "Total number of websites", Internet Live Stats.

13 *Idem*.

14 Peter Lyman, "Problem Statement: Why Archive the Web?", CLIR, consultado el 18 de mayo de 2023, https://www.clir.org/pubs/reports/pub106/web/#1.

y el 11 % de la información de redes sociales como X, antes Twitter, se borra al cabo de un año. La desaparición de la información publicada en la Web abre una brecha de conocimiento entre las actuales y futuras generaciones,[15] dado que la Web en el siglo XXI es el principal y más consultado medio de información con que cuenta la sociedad.

Además de los estudios antes mencionados, las personas que consultan Internet experimentan cómo las páginas web desaparecen constantemente. Se percatan de este hecho cuando buscan una página por la URL y reciben el mensaje "error 404 sitio no encontrado", lo que significa que la página no está más en el servidor.

Resulta paradójico que la expansión omnipresente de la Web se confronte con su naturaleza efímera y el riesgo constante de desaparición. Por ello, la preservación de las páginas web es un problema social que involucra la salvaguarda de una forma de nuestro patrimonio digital. El reconocimiento de su valoración como patrimonio digital es un tema de interés contemporáneo del que se han ocupado solo algunas comunidades de archivistas, bibliotecarios y profesionales de la información de instituciones de la memoria que reconocen la necesidad de preservar este tipo de información. En la mayoría de los casos hay otras prioridades. Por ejemplo, iniciar o concluir la digitalización de sendas colecciones impresas, sonoras, audiovisuales y fotográficas registradas en soportes analógicos. Instaurar archivos y repositorios digitales para la preservación de las colecciones digitales. Estabilizar y adecuar condiciones de conservación. Continuar y cuidar la consistencia de la catalogación. Diseñar servicios de información y atender a los usuarios. Divulgar y atraer mediante servicios innovadores a más usuarios.

Frente a estas tareas cotidianas que absorben el tiempo, la preservación de páginas web se observa como una tarea lejana aún de las bibliotecas y archivos. Esta situación se hace más difícil si se añade la falta de saberes, métodos y carencia de tecnologías para afrontar la salvaguarda de las páginas web. Lo cierto es que la salvaguarda de las páginas web

15 Miguel Costa, Daniel Gomes y Mário J. Silva, "La evolución del archivo web", *Revista Internacional de Bibliotecas Digitales* 18, (2017), https://doi.org/10.1007/s00799-016-0171-9.

es una tarea ineludible y cuanto más tiempo se postergue mayor será el volumen de material que se habrá perdido y con ello, se habrá borrado una parte de nuestro patrimonio digital.

1.3. Historia y evolución del Archivado Web

El Archivado Web es una práctica documental reciente. Su origen y desarrollo están asociados a la expansión y evolución tecnológica de la Web, así como a su naturaleza efímera y a su condición de riesgo de pérdida. De ahí que este modo de archivado, en casi tres décadas, ha evolucionado como resultado de diversas iniciativas de investigación y cooperación a nivel internacional creadas para preservar este tipo de documentos de origen digital.

Hasta 2017, se identificaron al menos 68 iniciativas[16] de Archivado Web impulsadas sobre todo por bibliotecas y archivos nacionales, en cuyo mandato de Depósito Legal se ha incorporado la preservación de la Web; así como por consorcios internacionales y compañías internacionales.

La evolución del Archivado Web se caracteriza por la innovación y desarrollo de *software* para automatizar los procesos documentales y por la instauración de formatos estandarizados para preservar este tipo de documentos. Se han desarrollado diversas herramientas para la recolección y acceso a este tipo de colecciones digitales.[17] Así como desarrollos para recuperar y poder consultar la información en su contexto de creación. A continuación, se presentan algunos de los hitos más significativos a través de los cuales se pueden observar la historia y evolución del Archivado Web.

En cuanto a la evolución de este tipo de materiales, debe señalarse que desde 1992 –año en que se publicó el primer sitio web– hasta aproximadamente el año 2000, las páginas web se basaron en texto. El uso de otros elementos gráficos era muy rudimentario. Sin embargo, en un

16 *Idem*.
17 *Idem*.

lustro, el desarrollo de herramientas fue rápido y para 2005, los sitios web ya incluían gráficos, búsquedas inteligentes, audio, video y contenidos animados, así como transmisiones en vivo.[18] En un muy corto periodo de tiempo la Web evolucionó y transitó del uso del lenguaje textual al de hipermedios. Y, además, se erigió como el canal dominante de comunicación y cimentó las bases para el desarrollo de plataformas y nuevos medios digitales.

Por otra parte, y a diferencia de otro tipo de publicaciones, las iniciativas para proteger este tipo de materiales captaron el interés de la comunidad internacional pocos años después de haber sido publicado el primer sitio. Es posible que, desde entonces, la fragilidad y carácter efímero de la Web se tornara en una de las preocupaciones de los creadores e impulsores de Internet.

Los primeros esfuerzos de acopio y salvaguarda de la Web se emprendieron en 1996 en Estados Unidos, Australia y Suecia, cinco años después de que Tim Berners-Lee publicara la primera página web en el CERN (European Organization for Nuclear Research). Por ello, de los primeros años de Internet solo se conservan algunas páginas que fueron resguardadas en *diskettes* y en discos compactos.[19]

Lideraron esta modalidad de preservación Internet Archive, PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) y el Swedish Kulturarw3, programa creado por la Royal Library of Sweden (Kungl Biblioteket) para la cosecha regular de imágenes instantáneas (*snapshots*) en el dominio nacional de ese país. Hasta el año 2000, la colección web de ese país nórdico ascendía a 65 millones de ítems, la mitad de los cuales eran documentos textuales, en formato HTML y texto plano.[20]

PANDORA fue un programa creado por la Biblioteca Nacional de Australia para que en colaboración con bibliotecas de ese país se recopilaran

18 Brown, *Archiving Websites*.

19 Costa, Gomes y Silva, "La evolución del archivo web".

20 Allan Arvidson, Krister Persson y Johan Mannerheim, "The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of 'complete' collection of web pages", 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 de agosto de 2000, https://archive.ifla.org/IV/ifla66/papers/154-157e.htm.

y conservaran publicaciones *online* relacionadas con Australia y los australianos. Con este programa se buscó atender a la Ley de Depósito Legal.[21] Para preservar el material recopilado y soportar los flujos de trabajo de la sección de archivo de la Biblioteca Nacional de Australia se puso en marcha desde 2001, PANDAS, sistema *open source* del archivo digital. En la actualidad el material recopilado se ofrece en acceso abierto en Trove Australian Web Archive.[22]

Internet Archive fue concebida como una biblioteca digital de sitios web de Internet. La iniciativa de salvaguarda tuvo sus orígenes en una empresa de catalogación creada por Brewster Kahle y Bruce Gilliant.[23] Internet Archive instauró el archivado de este tipo de materiales mediante la recolección o acopio remoto de páginas web. Para ello empleó Heritrix, considerado el primer *software* de recolección (*web crawling*) de sitios de manera remota que almacenó los recursos en un fichero ARC.

El ARC fue el primer formato de preservación de páginas web, después se consensuó y estableció el formato WARC. El método propuesto por Internet Archive prevalece en la actualidad,[24] aunque las tecnologías para desarrollar la cosecha de páginas web han evolucionado y automatizado este proceso a fin de hacerlo más sencillo y para poder cosechar un mayor volumen de contenidos. En 1999, Internet Archive ensanchó y complementó su ámbito de preservación de la Web a materiales que fueron digitalizados y nativos digitales como libros, documentos sonoros y audiovisuales.[25]

A más de dos décadas y media de distancia Internet Archive es, posiblemente, el archivo digital más grande del mundo que preserva una amplia gama de contenidos digitales, hasta mayo de 2023[26] comprendía:

21 "Pandora Overview", PANDORA Australia's Web Archive, consultado el 19 de mayo de 2023, https://pandora.nla.gov.au/overview.html.

22 "Websites", Trove, consultado el 19 de mayo de 2023, https://trove.nla.gov.au/search/category/websites?keyword=1991.

23 "About the Internet Archive", Internet Archive, consultado el 18 de mayo de 2023, https://archive.org/about/.

24 Brown, *Archiving Websites*.

25 *Idem*.

26 About the Internet Archive", Internet Archive.

- 735 000 millones de páginas web
- 41 millones de libros y textos
- 14.7 millones de grabaciones de audio (que incluyen 240 000 conciertos)
- 8.4 millones de video (que incluyen 2.4 millones de programas noticiosos y de televisión)
- 4.4 millones de imágenes
- 890 000 programas de *software*

Siguieron a las experiencias de PANDORA e Internet Archive diversas iniciativas de Archivado Web. El Nordic Web Archive (NWA) se creó en 1997 como una propuesta regional basada en la colaboración de las bibliotecas nacionales de Dinamarca, Finlandia, Islandia, Noruega y Suecia para intercambiar y desarrollar métodos y herramientas para el Archivado Web. Derivado de esta iniciativa se creó NWA Toolset, un paquete de *software* para acceder a documentos web archivados.[27]

Los miembros de la NWA impulsaron la creación del International Internet Preservation Coalition (IIPC) y desde 2003 se abocaron a desarrollar WERA (Web Archive Access),[28] un sistema de acceso que se basa en NWA Toolset y forma parte del desarrollo IIPC Toolkit.[29]

En enero de 1998, la Comisión Europea puso en marcha NEDLIB (Networked European Deposit Library) proyecto de investigación para garantizar que las publicaciones y los documentos electrónicos contemporáneos puedan utilizarse ahora y en el futuro. El proyecto fue liderado por la Biblioteca Nacional de Países Bajos, la Koninklijke Biblioteek, y se centró en desarrollar métodos de acopio de páginas web a fin de cumplir con la responsabilidad de Depósito Legal de las bibliotecas europeas.[30]

Por su parte, la Biblioteca del Congreso de los Estados Unidos emprendió la recolección y rastreo temático de sitios web desde el año

27 NWA, consultado el 18 de mayo de 2023, https://nwatoolset.sourceforge.net/.

28 "Introducción", IIPC, última publicación el 6 de noviembre de 2007, https://archive-access.sourceforge.net/projects/wera/.

29 Brown, *Archiving Websites*.

30 "Networked European deposit library", European Commission, consultado el 18 de mayo de 2023, https://cordis.europa.eu/project/id/LB5648.

2000, a través de MINERVA (Mapping the Internet Electronic Resources Virtual Archive) un proyecto piloto para proveer a la biblioteca de la capacidad de archivado web a gran escala.[31] Después se puso en marcha el Finland EVA Program, cuyas raíces se remontan a 1997, con el Swedish Kulturarw3, programa para la cosecha regular de imágenes instantáneas (*snapshots*) en el dominio nacional.

De acuerdo con las estadísticas proporcionadas por Wikipedia[32] hasta 2021 se han impulsado 97 iniciativas de Archivado Web. De estas ocho son mundiales, dos son producto de la colaboración entre países y el resto han sido iniciativas nacionales. Destaca Estados Unidos que ha promovido 29 y Alemania cinco. En África se documenta una iniciativa y en América Latina hay experiencias en la Biblioteca Nacional de Chile y en los archivos generales de Colombia y México. En el AGN de México en 2018 se organizó el Taller Preservación de Portales Web Institucionales con el uso de Webrecorder. La iniciativa pionera fue propuesta por la maestra Jo Ana Morfín y por el ingeniero Erick Cardoso en el AGN de México. Después de esta propuesta no se documenta que haya habido continuidad en el trabajo de archivado web.

Por otra parte, en México y desde la perspectiva de la investigación científica, en 2021 en el contexto de la pandemia, se puso en marcha el proyecto de investigación PAPIIT IT 400121, auspiciado por la Universidad Nacional Autónoma de México (UNAM), para el acopio y preservación de páginas web y redes sociales, en particular Twitter, sobre la pandemia de la COVID-19 en el país. Con ello, se inició la creación de colecciones digitales de la Web con fines de investigación científica.

De acuerdo con los datos anteriores se observa una brecha entre los países que desde hace tres décadas iniciaron proyectos de Archivado Web y de redes sociales. En tanto que hay países que desarrollan de forma sistemática y con herramientas cada vez más sofisticadas la preservación de la Web, subsisten regiones como África y América Latina, donde las propuestas para desarrollar este tipo de archivado son incipientes.

31 "Saving the World Wide Web", Library of Congress, consultado el 18 de mayo de 2023, https://www.digitalpreservation.gov/series/challenge/web_harvest_challenge.html.

32 "List of Web archiving initiatives", Wikipedia, última modificación el 10 de junio de 2023, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

De acuerdo con la información de Wikipedia,[33] 2009 fue el año en que se observó un mayor número de desarrollos porque se pusieron en marcha 13 iniciativas de Archivado Web. El establecimiento del International Internet Preservation Consortium (IIPC) en 2003 significó una etapa importante en la evolución del Archivado Web, porque se reunieron en este Consorcio algunos de los representantes de las principales iniciativas de Archivado Web que hasta entonces se habían creado a nivel mundial. El IIPC fue fundado con doce miembros. Participaron representantes de las bibliotecas nacionales de Australia, Canadá, Dinamarca, Finlandia, Francia, Islandia, Italia, Noruega y Suecia; la Biblioteca Británica y la del Congreso de los Estados Unidos e Internet Archive.[34] En la actualidad el Consorcio está compuesto por organizaciones de 35 países, entre las que se cuentan universidades, bibliotecas y archivos nacionales.

1.4. El contenido inabarcable de la Web

Hay quienes consideran que es imposible preservar la Web. Esta percepción coincide con un periodo de incertidumbre en torno a la preservación del patrimonio documental digital debido a la desaparición de ingentes cantidades de información como resultado de la obsolescencia de la tecnología y la falta de recursos para crear y mantener sistemas sustentables de preservación digital.

Quienes consideran que es imposible preservar la Web arguyen entre otros razonamientos: la magnitud, calidad y heterogeneidad de la información que se conserva por sí misma en Internet[35] y que, en muchos casos, se carece de las autorizaciones y los derechos de autor para preservar este material.

En cuanto a la magnitud de la Web conviene recordar, como se ha señalado con antelación, que el crecimiento de las páginas web es imparable. La dimensión de materiales que deben ser preservados es

33 *Idem*.
34 Brown, *Archiving Websites*.
35 Masanès, "Web Archiving".

tal que esta tarea no puede corresponder a una sola institución, país o persona. Debe ser una actividad colectiva de cooperación y colaboración entre las instituciones de la memoria y los especialistas que se han formado en el trabajo de campo. Así lo demuestran las iniciativas que desde hace casi tres décadas se han puesto en marcha para salvaguardar este tipo de materiales.

En relación con la calidad de la información digital debe señalarse que este es un argumento insostenible, porque aun cuando la información hipertextual de las páginas web no corresponde a los formatos de preservación digital estandarizados como son WAVE, MXF, PDF, entre otros –dado que en muchos casos se publican contenidos con compresión y pérdida–, la Web es un documento nativo digital que se codifica de un modo diferente a otro tipo de documentos. Desde hace más de una década se emplea el formato WARC como estándar y contenedor de información heterogénea producida en un lenguaje hipermedia.

Con respecto a la idea de que la Web se conserva a sí misma en Internet, esta noción es equivocada porque la información digital es un tipo de documento cuya posibilidad de ser borrada es alta. Los materiales digitales que pueblan la red son editados y no publicados. Es decir, que las publicaciones de Internet pueden ser modificadas o editadas de forma continua en tanto que los materiales impresos, sonoros o audiovisuales, por ejemplo, son producidos y se conserva la versión final de la obra. Esta perspectiva difiere de la idea de tratamiento de colecciones impresas, libros editados y publicaciones seriales que predominó en el siglo pasado. Dado que una vez publicados o transmitidos estos contenidos se entregaban para su preservación a diferentes instituciones de la memoria. En contraste, las páginas web pueden estar sometidas a una edición continua. Este es un rasgo que confiere vitalidad y oportunidad de actualización de la información, pero significa un problema para el profesional de la información que tiene a su cargo su salvaguarda.

Además, la noción generalizada de que estos materiales se resguardan en servidores y en los equipos de cómputo personales del creador y

con ello, se conservan de forma automática, es errónea. Los recursos en la Web son efímeros. El tiempo promedio de vida de las páginas web es de hasta dos años.[36] Las razones por las cuales los recursos desaparecen de la Web son diversas. Una de estas se refiere a la duración del dominio, de uno a tres años.[37] Esta duración afecta sobre todo a las páginas de instituciones u organizaciones pequeñas. Otra variable corresponde a la infraestructura tecnológica. Las páginas web requieren de tecnología que cada determinado tiempo entra en desuso y como consecuencia de este proceso se vuelve obsoleta. Este fenómeno afecta a los creadores de este tipo de publicaciones que a diferencia de los materiales impresos requieren de energía eléctrica, ancho de banda suficiente, dispositivos de almacenamiento y equipos que estén actualizados de forma permanente.

No obstante, en ciertas organizaciones públicas y privadas que disponen de forma sustentable de recursos tecnológicos, económicos y humanos; los cambios organizacionales pueden generar modificaciones significativas como son, por ejemplo, el cambio de la URL. De acuerdo con Berners Lee, en teoría no hay motivos para cambiar la URL, pero en la práctica sí existen millones de razones.[38]

Cuando se aduce que la preservación de la Web es una tarea imposible de llevar a cabo se sostiene que la cantidad de páginas es tal que esta deviene en una tarea titánica e imposible de llevar a cabo debido a las implicaciones del costo que tiene el almacenamiento de millones de páginas web. A las restricciones económicas y técnicas se suma otra de orden legal: la propiedad intelectual y los derechos de autor de los materiales publicados en la Web. Este tema que limita el acceso de sendas colecciones analógicas es aún más complejo en Internet porque contradice en muchos casos las iniciativas de acceso abierto al conocimiento científico.

Por otra parte, aunque en principio la Web es un espacio virtual abierto, las publicaciones pueden tener accesos limitados y requerir *login*

36 *Idem*.
37 *Idem*.
38 Berners-Lee *et al*., "The World-Wide Web".

y password. Un espacio protegido entonces no es público. Esta dualidad entre lo público y lo privado coexisten en la Web, condición que desde la perspectiva archivística conlleva a la reflexión en torno a saber si también los espacios privados deben ser preservados por instituciones de la memoria que se sostienen con recursos públicos.

Otra discusión reciente en relación con la preservación de la Web involucra el derecho al olvido como el derecho que tiene cualquier persona para impedir la difusión de información y datos personales obsoletos o que ya no tienen relevancia e interés público, toda vez que pueden afectar su privacidad e identidad.

Bajo las consideraciones anteriores, la preservación de la Web parece una tarea titánica inabarcable. Sin embargo, la desaparición de un fragmento de la información publicada en la Web significa la pérdida de una parte de la memoria de la era digital.

II. El archivado de las redes sociales

2.1 ¿Qué son las redes sociales?

Acercamientos teóricos en los estudios de redes sociales no coinciden en una definición del concepto.[39] La antropóloga social y cultural Isabel Ponce propone una definición en la que destaca una serie de características principales de las redes sociales:

> (...) podemos definir las redes sociales *on-line* como estructuras sociales compuestas por un grupo de personas que comparten un interés común, relación o actividad a través de Internet, donde tienen lugar los encuentros sociales y se muestran las preferencias de consumo de información mediante la comunicación en tiempo real, aunque también puede darse la comunicación diferida en el tiempo, como en el caso de los foros. No sólo nos relacionamos y compartimos con los demás, sino que además, exponemos abiertamente y en tiempo real nuestros gustos y tendencias, expresando la propia identidad.[40]

En esta definición podemos identificar un énfasis en el uso colectivo de un espacio virtual que permite la comunicación entre personas en

39 Caleb T. Carr y Rebecca A. Hayes, "Social media: defining, developing, and divining", *Atlantic Journal of Communication* 23, n.° 1 (2015), 46.

40 Isabel Ponce, "Monográfico: Redes Sociales – Definición de redes sociales", Observatorio Tecnológico, publicado el 17 de abril de 2012, http://recursostic.educacion.es/observatorio/web/en/internet/web-20/1043-redes-sociales?start=1.

tiempo real. Estos espacios, a su vez, permiten la interacción dinámica entre diversos tipos de público sin importar elementos de espacio y distancia. En este contexto, las plataformas de redes sociales juegan un papel como medio, por el cual se pueden llevar a cabo estas dinámicas de comunicación, caracterizadas por la interactividad y la participación global.[41]

Ariel y Avidar[42] realizan un ejercicio de introspección sobre la conceptualización de las redes sociales, profundizando a partir de las diversas definiciones que se le han dado al concepto. En específico, presentan un modelo teórico para comprender lo social en las redes sociales basado en las relaciones entre tres conceptos vinculados a los ambientes de las redes sociales: información, interactividad y sociabilidad. En el contexto de las redes sociales, el concepto de información puntualiza tanto la práctica de transmisión como la de dar significado a dicha información por parte de los usuarios, permitiendo a su vez que estos sean co-creadores de significado.[43] Estos procesos de transmisión de información y de dar significado se ven reflejados en la interactividad.[44] Ariel y Avidar advierten que la interactividad no es un elemento inherente de las tecnologías, sino que tecnologías como las desarrolladas para redes sociales[45] permiten la interactividad. En ese marco, los atributos tecnológicos de las redes sociales no son suficientes para concluir que estas son inherentemente sociales. Son las actividades, los tipos de interacciones y los niveles de participación de los usuarios los elementos que definen la sociabilidad de las redes sociales. En resumen, en los procesos de comunicación a través de las redes sociales la información es su unidad fundamental, pero son los usuarios quienes definen a través de sus actividades los niveles de interactividad y sociabilidad de tales redes.[46] Vinculando este modelo

41 Zahira Moreno Freites y Gertrudis Ziritt Trejo, "Redes Sociales como canales de digi-impacto en la participación ciudadana", *Utopía y Praxis Latinoamericana. Revista Internacional de Filosofía Iberoamericana y Teoría Social* 24, n.° 3 (2019): 30.

42 Yaron Ariel y Ruth Avidar, "Information, Interactivity and Social Media", *Atlantic Journal of Communication* 23, n.° 1 (2015).

43 *Ibid*., 21.

44 *Ibid*., 24.

45 *Ibid*., 26.

46 *Ibid*., 28.

teórico-conceptual a los procesos de Archivado Web argumentamos que uno de los propósitos principales del archivado de redes sociales, y a su vez uno de los más desafiantes, es buscar no solo la preservación de la información que se transmite, sino además la interactividad y sociabilidad que se lleva a cabo en estos espacios.

Por su parte, José Van Dijck presenta un análisis sobre las plataformas de redes sociales y lo que llama la cultura de la conectividad para comprender estas plataformas como constructos tecnoculturales donde se vinculan la tecnología, los usuarios y el contenido y como estructuras socioeconómicas donde se reflejan elementos de propiedad, gobierno y modelos de negocio.[47] Estos dos niveles (constructos tecnoculturales y estructuras socioeconómicas) y seis elementos constitutivos (tecnología, usuarios, contenido, propiedad, gobierno y modelos de negocio) representan a cada plataforma como un microsistema desarrollado a partir de la interdependencia de estos niveles y elementos, así como del contexto social, político y económico del ecosistema de las redes sociales y los medios de comunicación tradicionales.[48] Al respecto explica Van Dijck:

> Lo más característico del ecosistema es la *interdependencia* y la *interoperabilidad* de las plataformas. Analizar los seis elementos constitutivos de cada una de las plataformas estudiadas permitirá revelar patrones específicos de funcionamiento del ecosistema. Los botones de "compartir", "seguir", "convertir en tendencia" y "marcar como favorito" tienen cada uno características diferenciales, pero poseen una lógica común: la implementación ubicua del botón de una plataforma competidora no sólo indica un alineamiento tecnológico, sino también una maniobra estratégica que procura estimular el tráfico de usuarios e infiltrarse en sus rutinas. Por ejemplo, la integración de la categoría de "trending topic" de Twitter en otras plataformas, y también en medios tradicionales como las cadenas de noticias y entretenimiento, ejerce una profunda influencia en las prácticas profesionales de los periodistas y en los hábitos de los usuarios. La fusión tecnológica entre

47 José Van Dijck, *La cultura de la conectividad: una historia crítica de las redes sociales* (Buenos Aires: Siglo XXI Editores, 2016), 33.

48 *Ibid*., 44.

> las distintas plataformas y la influencia conjunta sobre los usuarios y el contenido indican que difícilmente sea posible estudiar a los microsistemas por separado.[49]

Una de las características principales atribuida a las redes sociales es ser espacios de expresión y democratización de la información. Tal vez, el ejemplo más común que se presenta es el de la Primavera Árabe, la cual tuvo sus raíces de protestas masivas en contra de las condiciones de vida y corrupción en Túnez a raíz del acto de Mohamed Bouazizi, quien se inmoló en la calle al ser despojado por la policía de su mercancía y dinero.

Las protestas causaron la salida del presidente Ben Ali. Estos tipos de protestas se extendieron a otros países de la región como Egipto, Libia y Yemen, y en todos las plataformas de redes sociales jugaron un papel protagónico.[50] En Estados Unidos el movimiento *Black Lives Matter* –que protesta contra los asesinatos y abusos a afroamericanos por parte de la policía– utiliza las redes sociales como uno de sus medios principales de comunicación, construcción de redes y activismo digital.[51] Mientras que en México, las redes sociales jugaron un papel importante en el movimiento *Yo Soy 132*, el cual surgió luego de la visita del entonces gobernador del Estado de México y candidato presidencial por el Partido Revolucionario Institucional (PRI), Enrique Peña Nieto, a la Universidad Iberoamericana, en donde fue cuestionado por estudiantes debido a los actos de represión cometidos por los aparatos de seguridad en San Salvador Atenco contra manifestantes que protestaban por la construcción del aeropuerto internacional en Texcoco.[52]

---

49 *Idem*.

50 Miguel de Aguilera y Andreu Casero Ripollés, "¿Tecnologías para la transformación? Los medios sociales ante el cambio político y social. Presentación". *Icono 14, Revista de Comunicación y Tecnologías Emergentes* 16, n.º 1 (2018): 3.

51 Marcia Mundt, Karen Ross y Charla M. Burnett, "Scaling Social Movements Through Social Media: The Case of Black Lives Matter", *Social Media + Society* 4, n.º 4 (October-December 2018).

52 "A 10 años del origen del #YoSoy132 en la IBERO, su legado sigue vigente", Ibero Ciudad de México, publicado el 9 de mayo de 2022, https://ibero.mx/prensa/10-anos-del-origen-del-yosoy132-en-la-ibero-su-legado-sigue-vigente; Guiomar Rovira Sancho, "El #YoSoy 132 mexicano: la aparición (inesperada) de una red activista/The Mexican# YoSoy 132: the (unexpected) emergence of a activist network", *Revista CIDOB d'Afers Internacionals*, n.º 105 (2014), https://www.cidob.org/es/articulos/

Sin embargo, esta visión de las redes sociales como espacios democráticos de expresión ha sido cuestionada, esto a partir del análisis sobre el poder de los algoritmos en la manera en que rigen el funcionamiento de las plataformas de redes sociales. Como explica el profesor de la Universidad de California, en Berkeley, Stuart Russell, en una entrevista a la BBC Brasil, el objetivo principal de los algoritmos de las plataformas de redes sociales es recopilar datos de cada usuario para identificar patrones de comportamiento en la red social y de ahí proporcionar contenido que se adapte a esos patrones para mantener al usuario conectado y activo, lo cual a su vez abre las puertas a la manipulación de las personas.[53] A esto añadimos que el modelo de negocios de las plataformas de redes sociales se basa en el uso de los datos de los usuarios para generar ingresos. Shoshana Zuboff presenta el concepto de «capitalismo de la vigilancia» para representar este modelo, el cual ha estado rigiendo estructuras económicas en el siglo XXI.[54] El objetivo principal del capitalismo de vigilancia es la extracción y acumulación de datos que permitan analizar el comportamiento de los usuarios para a su vez intentar predecir y/o modificar comportamientos que redunden en ingresos para las empresas.[55]

Esta arquitectura de extracción ha sido también utilizada para incidir en la discusión pública sobre temas sociales y políticos, donde la desinformación ha jugado un papel predominante. El caso de Cambridge Analytica y la elección presidencial de 2016 en Estados Unidos ilustra esta práctica. Cambridge Analytica era una empresa consultora en temas políticos. En marzo de 2018 medios como el New York Times y el Observer reportaron que la empresa obtuvo datos de 50 millones de usuarios de Facebook, los cuales fueron utilizados para implementar algoritmos para identificar perfiles de votantes estadounidenses y lanzar campañas políticas personalizadas a favor de la candidatura de Donald

revista_cidob_d_afers_internacionals/105/el_yosoy132_mexicano_la_aparicion_inesperada_de_una_red_activista.

53 Paula Adamo Idoeta, "Por qué los algoritmos de las redes sociales son cada vez más peligrosos", BBC News Mundo, 12 de octubre de 2021, https://www.bbc.com/mundo/noticias-58874170.

54 Shoshana Zuboff, *The Age of Surveillance Capitalism* (New York: PublicAffairs, 2018).

55 *Ibid*., 131.

Trump.[56] Como explica Zuboff, más allá del impacto político que tuvo esta acción en los resultados de la elección presidencial, el nivel de planificación y la ambición de Cambridge Analytica sobre los potenciales resultados de este modelo de extracción para influir en los discursos públicos "son testamento del papel esencial de la interpretación profunda en la predicción y modificación de comportamientos, siempre en busca de la certeza".[57]

2.2 Historia y evolución de las redes sociales

En este apartado presentamos un panorama general sobre la evolución de las plataformas de redes sociales desde finales del siglo xx. El propósito es ilustrar cambios significativos en las funcionalidades e interactividad de las redes sociales en los últimos 30 años.

En mayo de 1996, Andrew Weineich fundó Six Degrees, un sitio web con las características que se ven actualmente en plataformas de redes sociales: crear perfiles y conectarse a otros usuarios. El sitio sixdegrees.com fue lanzado un año después y es considerado como la primera plataforma de redes sociales.[58] Su diseño y funcionalidad fue inspirada por la teoría de los seis grados, la cual argumenta que cada persona está conectada a otra por no más de seis vínculos de relación.[59] Usuarios de la plataforma podían crear un perfil y ver los de sus amigos, enviar mensajes y publicar contenidos en los foros de personas vinculadas a sus primeros tres grados de conexión.[60] Tres años después, en el 2000, la plataforma cerró operaciones, principalmente porque para finales de

56 Matthew Rosenberg, Nicholas Confessore y Carole Cadwalladr, "How Trump Consultants Exploited the Facebook Data of Millions", *The New York Times,* 17 de marzo de 2018; "The Cambridge Analytica Files", *The Guardian*, consultado el 7 de julio de 2023, https://www.theguardian.com/news/series/cambridge-analytica-files.

57 Zuboff, *The age of*, 282.

58 Chenda Ngak, "Then and now: a history of social networking sites", *CBS News*, 6 de julio de 2011, https://www.cbsnews.com/pictures/then-and-now-a-history-of-social-networking-sites/.

59 "Six Degrees: cómo fue y quién creó la primera red social de internet, inspirada por la teoría de los «seis grados»", *BBC News Mundo*, 8 de junio de 2019, https://www.bbc.com/mundo/noticias-48558989.

60 Adeola Adeyemo, "History & importance of the first social media site - Six Degrees", Adeola, publicado en 2018, https://adeolawrites.com/history-importance-of-the-first-social-media-site-six-degrees/.

los 90 muchos usuarios de Internet aún no tenían redes de amigos en la Web, además de que tampoco contaban con una conexión más rápida de Internet.[61]

Entre las plataformas que surgieron a inicios del siglo xxi podemos destacar Friendster, lanzada en el 2002 como una red social dirigida a ayudar con la creación de conexiones entre amigos de amigos y de interés romántico para competir con el sitio web de citas Match.com, y que en marzo de 2003 contaba con 300 000 usuarios.[62] La popularidad de Friendster impactó negativamente la plataforma debido a que la infraestructura tecnológica no contaba con las condiciones para manejar su crecimiento.[63] Esta realidad y el surgimiento de Facebook causó el colapso de Friendster en Estados Unidos, aunque se mantuvo a flote unos años adicionales por su amplio uso en el sureste de Asia.[64]

Un año después del lanzamiento de Friendster, Chris DeWolfe y Tom Anderson lanzaron *MySpace*, red social que se convirtió en la plataforma más visitada en el mundo de 2005 a 2008.[65] Entre las funcionalidades de *MySpace* se encontraban los foros, chats y espacios para compartir videos y música. Esto último fue una de las funcionalidades más fuertes de la plataforma, donde artistas y bandas independientes compartían su música con los usuarios. La banda de rock *Artic Monkeys* fue una de ellas.[66] Para septiembre de 2005, *MySpace* contaba con 27 millones de usuarios.[67] Ya para el año 2008, la plataforma comenzó a perder popularidad ante la competencia de otras redes sociales, particularmente *Facebook* que para ese año logró superar a *MySpace* en

61 "Six Degrees: cómo fue", *BBC News Mundo*.

62 Danah M. Boyd y Nicole B. Ellison, "Social Network Sites: Definition, History and Scholarship", *Journal of Computer-Mediated Communication* 13, n.° 1 (1 de octubre de 2007): 215, https://doi.org/10.1111/j.1083-6101.2007.00393.x.

63 *Idem*.

64 Robert McMillan, "The Friendster Autopsy: How a Social Network Dies", *Wired*, 27 de febrero de 2013, https://www.wired.com/2013/02/friendster-autopsy/.

65 Lori Kozlowski, "New life: How MySpace Spawned a Start-Up Ecosystem", *Forbes*, 15 de mayo de 2012, https://www.forbes.com/sites/lorikozlowski/2012/05/15/how-myspace-spawned-a-start-up-ecosystem/?sh=2dee0c6040ba.

66 *Idem*.

67 Nicholas Jackson y Alexis C. Madrigal, "The Rise and Fall of MySpace", *The Atlantic*, 12 de enero de 2011, https://www.theatlantic.com/technology/archive/2011/01/the-rise-and-fall-of-myspace/69444/.

número de usuarios en Estados Unidos.[68] Actualmente *MySpace* se presenta como un sitio web enfocado en temas relacionados con la música, en el 2019 tuvo aproximadamente siete millones de visitas.[69] En un evento que ejemplifica los desafíos de la preservación digital, en marzo de 2018 *MySpace* anunció que posiblemente perdió su contenido de los pasados doce años, aludiendo a una falla en el proceso de migración a un servidor.[70]

En este breve relato histórico de la evolución de las plataformas de redes sociales se logran identificar algunos de los impactos que tuvo en *Friendster* y *MySpace* la llegada de *Facebook*. Fundado en el 2004 por Mark Zuckerberg, Eduardo Saverin, Dustin Moskovitz y Chris Hughes en la Universidad de Harvard, su consolidación en el universo de las plataformas de redes sociales se ha visto a lo largo de los últimos 15 años. En marzo de 2012, *Facebook* contaba con 835 millones de usuarios a nivel mundial.[71] Para marzo de 2023, el número de usuarios activos por mes fue de 2.98 billones en todo el mundo.[72] La empresa de Zuckerberg, que en octubre de 2021 cambió el nombre a *Meta*, también administra las redes sociales Instagram y WhatsApp, adquiridas en 2012 y 2014, respectivamente. En julio de 2023, *Meta* lanzó la red social Threads, una plataforma con funcionalidades muy similares a las de Twitter.

68 Nate Williams, "The Real Reason MySpace Failed Spectacularly", History-Computer, publicado el 13 de diciembre de 2022, https://history-computer.com/the-real-reason-my-space-failed-spectacularly/#:~:text=At%20its%20prime%2C%20MySpace%20was,mass%20shift%20to%20new%20platforms.

69 Elise Moreau, "Is MySpace Dead? The troubled social network's struggle to make a real comeback", Lifewire Tech for Humans, actualizado el 21 de enero de 2022, https://www.lifewire.com/is-myspace-dead-3486012#:~:text=The%20Current%20State%20of%20Myspace,-Officially%2C%20however%2C%20Myspace&text=If%20you%20go%20to%20myspace,over%207%20million%20monthly%20visits.

70 "MySpace: el error en un servidor que hizo perder a la red social 12 años de música almacenada", *BBC News Mundo*, 18 de marzo de 2019, https://www.bbc.com/mundo/noticias-47612172.

71 Van Dijck, *La cultura*, 50.

72 Stacy Jo Dixon, "Facebook: quarterly number of MAU (monthly active users) worldwide 2008-2023", Statista, publicado el 9 de mayo de 2023, https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/#:~:text=With%20roughly%202.98%20billion%20monthly,used%20online%20social%20network%20worldwide.

Palabras como "compartir", "me gusta" y "hacer amigos" se vinculan con las funcionalidades principales de Facebook a través de su evolución. Al respecto, Van Dijck explica lo siguiente:

> La ideología de "compartir", sostenida por Facebook, estableció en buena medida el estándar para las demás plataformas y para el ecosistema en su totalidad. Debido a su posición de liderazgo indiscutido dentro del segmento de los sitios de red social, sus prácticas influyeron de manera sustantiva en las normas sociales y culturales que apuntalan valores legales tales como la privacidad y el control sobre la información.[73]

Las prácticas de *Meta* en relación con la privacidad y el control sobre la información han generado críticas y conflictos legales en Estados Unidos y particularmente en países europeos. Una de las críticas principales, de las cuales Meta no ha sido el único, es cómo la empresa extrae y utiliza datos relacionados con el comportamiento de usuarios y no usuarios de sus plataformas para generar publicidad dirigida ("ad targeting"). Como consecuencia del escándalo de Cambridge Analytica, Zuckerberg fue citado a testificar frente al Congreso de los Estados Unidos en abril de 2018, donde fue duramente cuestionado por las prácticas de minería de datos de Facebook.[74] Tres años antes, la Comisión de Privacidad de Bélgica ordenó a Facebook descontinuar la extracción de datos de no-usuarios de la red social sin su consentimiento.[75] Y en 2019, autoridades antimonopolio en Alemania ordenaron a Facebook no extraer datos de usuarios de otras plataformas como Instagram para sus prácticas de minería de datos sin el consentimiento de los usuarios.[76]

---

73 Van Dijck, *La cultura,* 50-51.

74 Natasha Singer, "What You Don't Know About How Facebook Uses Your Data", *The New York Times*, 11 de abril de 2018, https://www.nytimes.com/2018/04/11/technology/facebook-privacy-hearings.html.

75 *Idem*.

76 David Rising y Barbara Ortutay, "Germany to Facebook: Stop forcing users to share their data", *The Associated Press*, 7 de febrero de 2019, https://apnews.com/article/ap-top-news-facebook-privacy-scandal--social-platforms-germany-north-america-04440c1ca08b4caf9da2f6e9bf0038d7.

Mientras que la filosofía principal de Facebook ha rondado alrededor del concepto de "compartir", Twitter consolidó en el ecosistema de las redes sociales la práctica de *microblogging*, una manera de comunicación por medio de mensajes cortos. Tras su lanzamiento en 2007 se le apodó el "SMS de internet" y se le consideraba como un intermedio entre prácticas de comunicación como el correo electrónico, los mensajes de texto y las llamadas telefónicas al tener un límite de 140 caracteres por tuit.[77] En 2014 ya contaba con casi 500 millones de usuarios registrados en todo el mundo y 88 millones de usuarios activos por mes,[78] en diciembre de 2022 el número de usuarios activos era de 388.4 millones.[79] Entre las funcionalidades que caracterizan Twitter y que influyeron en otras plataformas de redes sociales se encuentran las etiquetas (o *hashtags*), implementadas en el 2007.[80] Por medio de las etiquetas los usuarios articulan palabras o frases precedidas por el signo de número (#). A finales de 2008, Twitter incorpora las "tendencias" (*trending topics*), mostrando a los usuarios una lista de temas que según el algoritmo estaban generando mayor conversación.[81] A través del tiempo, Twitter se convirtió en una de las plataformas de redes sociales más utilizadas por medios de comunicación y periodistas para compartir noticias. Además, y como se explica en la sección anterior, ha sido utilizada como medio de comunicación por movimientos sociales. Sin embargo, esta red social de *microblogging* se ha convertido en espacio para la diseminación de desinformación, como por ejemplo, en relación con la COVID-19.[82] El análisis histórico de Twitter realizado por José Van Dijck, publicado antes de la pandemia, ya identificaba esta problemática de la red social:

77 Van Dijck, *La cultura*, 75.

78 *Ibid*., 74.

79 Statista Research Department, "Number of Twitter users worldwide from 2019 to 2024", Statista, publicado el 14 de diciembre de 2022, https://www.statista.com/statistics/303681/twitter-users-worldwide/.

80 Belle Beth Cooper, "The Surprising History of Twitter's Hashtag Origin and 4 Ways to Get the Most out of Them", Buffer, publicado el 24 de septiembre de 2013, https://buffer.com/resources/a-concise-history-of-twitter-hashtags-and-how-you-should-use-them-properly/.

81 Van Dijck, *La cultura*, 76.

82 David García-Marín y Marta Merino-Ortego, "Desinformación anticientífica sobre la COVID-19 difundida en Twitter en Hispanoamérica", *Cuadernos.info* 52 (mayo 2022).

> A pesar de la imagen de Twitter como una "asamblea municipal" *online* para la comunicación colectiva –un mero amplificador de voces individuales y opiniones colectivas–, el sitio ha comenzado a mostrarse cada vez más como un potente instrumento para fomentar ideas y manipular opiniones.[83]

Twitter también es un buen caso para reflexionar sobre cuán social es una plataforma de red social. Retomando el análisis de Ariel y Avidar sobre los niveles de interactividad y sociabilidad de las redes sociales, las estadísticas muestran que un porcentaje pequeño del universo de usuarios de Twitter llevan a cabo la mayoría de las conversaciones. Un estudio del Pew Research Center encontró que en Estados Unidos el 25 % de los usuarios generan el 97 % de los tuits.[84]

En octubre de 2022, Elon Musk finalizó la adquisición de Twitter por 44 billones de dólares y comenzó a introducir cambios sustanciales en la operatividad de la red social. Esto incluyó la eliminación del equipo de trabajo a cargo de la moderación de contenidos y el establecimiento de un plan que requiere pagar ocho dólares mensuales para mantener la palomita de verificación azul y solicitar el pago de una mensualidad para tener acceso a la Interfaz de Programación de Aplicaciones (API por sus siglas en inglés), impactando en el acceso de investigadores y curadores digitales, entre otras poblaciones. Además, en julio de 2023, Musk cambió el nombre de la plataforma a X.[85]

¿Cuál es el panorama actual sobre uso de plataformas de redes sociales en México? Según datos publicados por Statista, en el 2022 el país contaba con aproximadamente 98 millones de usuarios de redes sociales.[86] En cuanto a las plataformas de redes sociales más utilizadas,

83 Van Dijck, *La cultura*, 79.

84 Meltem Odabas, "10 facts about Americans and Twitter", Pew Research Center, publicado el 5 de mayo de 2022, https://www.pewresearch.org/short-reads/2022/05/05/10-facts-about-americans-and-twitter/.

85 Chantal Da Silva, "Twitter rebrands to 'X' as Elon Musk Loses iconic bird logo", *NBC News*, 24 de julio de 2023, https://www.nbcnews.com/news/us-news/twitter-rebrands-x-elon-musk-loses-iconic-bird-logo-rcna95880#.

86 Statista Research Department, "Número de usuarios de redes sociales en México de 2017 a 2027", Statista, publicado el 27 de marzo de 2023, https://web.archive.org/web/20230523033503/https://es.statista.com/estadisticas/1141228/numero-de-usuarios-de-redes-sociales-mexico/.

la lista la lideraba Facebook con un 92.9 % de los usuarios, seguida muy de cerca por WhatsApp (92.2 %). Por su parte, Facebook, Messenger e Instagram fueron utilizadas por aproximadamente el 80 % de los usuarios, mientras que TikTok contaba con cerca del 74 % de la población de usuarios. Twitter se ubica en un sexto lugar con 53.7 % de usuarios.[87] Según este análisis, el grupo de edad con el mayor porcentaje de usuarios de redes sociales es de 25 a 34.[88]

En agosto de 2022, en el marco del Día Internacional de la Juventud, el Instituto Nacional de Estadística y Geografía (INEGI) publicó estadísticas sobre algunas características de la población de 12 a 29 años, lo que incluyó el uso de redes sociales. Según el INEGI, 33.9 millones de jóvenes interactuaron en redes sociales. Las plataformas más utilizadas por los jóvenes fueron: WhatsApp (92.3 %), Facebook (90.6 %), Instagram (50.3 %), Facebook Messenger (43.4 %) y YouTube (36 %).[89]

Ciertamente la historia de la evolución de las redes sociales desde finales del siglo XX es mucho más compleja que el resumen que se presenta en esta sección. Sin embargo, con este panorama general podemos identificar cómo las plataformas de redes sociales se han insertado en el día a día e impactado en diversas experiencias sociales. Estos efectos, a su vez, han generado reflexiones sobre los contenidos de las redes sociales como potenciales fuentes de investigación histórica y para documentar las sociedades. En el siguiente apartado exploramos esas consideraciones.

2.3 El archivado de las redes sociales

¿Cuál es la importancia de preservar contenidos de redes sociales? Un argumento gira alrededor de que ante la proliferación de las redes so-

87 Statista Research Department, "Porcentaje de usuarios por red social en México en 2022", Statista, publicado el 28 de marzo de 2023, https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/.

88 Statista Research Department, "Las redes sociales en México – Datos estadísticos", Statista, publicado el 27 de marzo de 2023, https://es.statista.com/temas/7392/las-redes-sociales-en-mexico/#topicOverview.

89 Instituto Nacional de Estadística y Geografía, "Estadísticas a propósito del Día Internacional de la Juventud" (comunicado de prensa 436/22, 10 de agosto de 2022), https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_Juventud22.pdf.

ciales se genera a su vez un caudal significativo de información que permite estudiar fenómenos sociales.[90] Un estudio de Hemhill, Hedstrom y Leonard encontró que algunos investigadores utilizaron contenidos de Twitter para estudiar varios ámbitos, incluyendo el análisis de redes en el contexto de temas políticos y eventos.[91]

En diversos espacios académicos se han establecido laboratorios enfocados en realizar análisis de datos de redes sociales para estudiar tendencias, opiniones, etcétera. Por ejemplo, la Universidad Veracruzana cuenta desde el 2017 con el Laboratorio para el Análisis de Información Generada a través de Redes Sociales en Internet (LARSI), la cual se enfoca en realizar estudios de opinión a través del análisis de datos de redes sociales (https://www.uv.mx/larsi/).

La proliferación de información a través de las redes sociales no es la única razón, y tal vez no la más importante, para argumentar en favor de la preservación de estos contenidos. Desde el contexto de la historia archivística lo podemos colocar como un elemento adicional dentro de la evolución de teorías y prácticas archivísticas, particularmente a partir del siglo XIX, periodo en el cual se solidifica la archivística como ciencia.[92] Desde el contexto digital, en la década de los 90 surge un fortalecimiento en la investigación archivística para atender los desafíos sobre la gestión y preservación de documentos electrónicos. En 1991, Margaret Hedstrom hizo un llamado para desarrollar estrategias metodológicas y atender las particularidades de la preservación de documentos electrónicos. Hedstrom reconoce que el contexto de documentos electrónicos plantea reformulaciones a teorías y prácticas archivísticas, lo cual abriría

90 Joel Antonio Blanco-Rivera, "La archivología en el contexto de la sociedad interconectada por redes", *Revista Interamericana de Bibliotecología* 42, n.º 3 (2019), https://doi.org/10.17533/udea.rib.v42n3a02; Ian Milligan, "La historia en la era de la abundancia: archivos web e investigación histórica", *Historia y Memoria*, n.º especial 10 años (2020).

91 Libby Hemphill, Margaret L. Hedstrom y Susan Hautaniemi Leonard, "Saving social media data: understanding data management practices among social media researchers and their implications for archives". *Journal of the Association for Information Science and Technology* 72, n.º 1 (2021): 102.

92 Leomar José Montilla Peña y Mayra M. Mena Mujica, "Estado de desarrollo de la archivística clásica hasta los años 30 del siglo XX: Tres manuales archivísticos de trascendencia universal", *Biblios* 52 (2013).

las puertas a que archivistas se insertaran en trabajos interdisciplinarios para contribuir con el diseño de sistemas de información que tomen en cuenta elementos de preservación. Sobre esto, Hedstrom menciona:

> Los archivistas se encuentran en una posición única para contribuir al diseño de los sistemas de información por su perspectiva singular sobre la relación entre la misión y la estructura de una organización, su necesidad de producir documentos, sus flujos de información, y estructuras documentales.[93]

Ya dentro del contexto de la archivología en el siglo XXI, autoras como Gilliland[94] y Sheffield[95] hacen un llamado muy similar al realizado por Hedstrom a inicios de los 90, insertando reflexiones sobre la relación archivos, poder y sociedad.

En cuanto a los archivos de redes sociales, las investigaciones de Amelia Acker y Adam Kriesberg son fundamentales para comprender y analizar las implicaciones del archivado de redes sociales en teorías y prácticas archivísticas. Para propósitos de este texto enfocaremos dos aspectos fundamentales sobre el archivado de redes sociales que Acker, Kriesberg y otros autores han examinado. El primero está relacionado con la contextualización. En "Tweets may be archived: civic engagement, digital preservation and Obama White House social media data", Acker y Kriesberg explican que la naturaleza misma de las redes sociales presenta desafíos de prácticas de preservación digital, las cuales confrontan las técnicas comunes de archivado de la Web, esto principalmente, porque la naturaleza misma de los datos estructurados que dan forma a los contenidos de redes sociales conforman flujos de información y metadatos complejos.[96]

93 Margaret Hedstrom, "Understanding Electronic Incunabula: A Framework for Research on Electronic Records", *The American Archivist* 54, n.º 3 (1 de julio de 1991): 338.

94 Anne J. Gilliland, "Reconceptualizing Records, the Archive and Archival Roles and Requirements in a Networked Society", *Knygotyra* 63 (2014). DOI: 10.15388/kn.v63i0.4011.

95 Rebecka Taves Sheffield, "Facebook Live as a Recordmaking Technology", *Archivaria* 85 (2018).

96 Amelia Acker y Adam Kriesberg, "Tweets may be archived: Civic engagement, digital preservation and Obama White House social media data", *Proceedings of the Association for Information Science and Technology* 54, n.º 1 (2017): 4.

Estas complejidades reflejan, desde el punto de vista de la preservación digital, los desafíos para contextualizar la data que se está recopilando y preservando. Acker y Kriesberg aplicaron métodos de análisis de datos y de carácter forense digital para analizar los archivos de redes sociales creados desde las cuentas del presidente Barack Obama, específicamente sus cuentas de Facebook, Twitter y Vine. Este análisis generó observaciones importantes sobre temas de procedencia y metadatos, fundamentales para la contextualización. En el caso del archivo de Facebook, por ejemplo, Acker y Kriesberg encontraron muchas limitaciones para realizar búsquedas, particularmente con las fotos, las cuales no contienen metadatos básicos como título y año. El tema de los metadatos también fue identificado en el caso del archivo de tuits, donde se pierden metadatos, particularmente relacionados a retuits y *likes*, cuando se realiza el proceso de extracción de los datos de la plataforma Twitter. En este sentido, Acker y Kriesberg explican que uno de los desafíos principales para la preservación de tuits es identificar su procedencia a partir del análisis de métricas de veracidad e interactividad debido al contexto que se pierde al realizar la extracción de datos.

> Los datos de las redes sociales pierden contexto importante una vez que se extraen de su plataforma nativa, convirtiéndolos en una captura de ese momento en el tiempo. Los metadatos y otras características pueden cambiar debido a los cambios en las reglas de las APIs, rediseño de las plataformas y otras decisiones que no se toman considerando sus implicaciones en la investigación académica, pero que afectan la misma. Si la precisión de los metadatos de interactividad no es confiable el valor de Twitter como un activo para la investigación social disminuirá, a menos que los investigadores desarrollen nuevos conceptos para considerar las lagunas en estos materiales.[97]

Tres años después, Acker y Kriesberg realizaron un nuevo análisis sobre las implicaciones de los archivos de redes sociales, específicamente en

97 *Ibid*., 7.

relación con la preservación de contenidos que son manejados/accedidos/manipulados a través de las API, lo cual conforma la base de las plataformas de redes sociales para almacenar, compartir y gestionar datos.[98] Acker y Kriesberg analizan las API desde la perspectiva teórica de archivos post-custodiales, una noción que surge a partir de la década de los 80, lo que implica que productores de documentos mantengan su custodia y donde el archivista asume un rol de colaborador que permite dar seguimiento a procesos de gestión y preservación de documentos. Esta noción estaba enfocada en la gestión de documentos (particularmente electrónicos), pero ha evolucionado y ha sido incorporada en proyectos de archivos comunitarios donde se aplican prácticas en las que la propia comunidad mantiene la custodia de sus archivos, aplicando estrategias digitales para la difusión de las narrativas de las comunidades, dando así mayor agencia a las propias comunidades sobre sus archivos.[99]

Las API pueden ser situadas en este contexto post-custodial. Sin embargo, explican Acker y Kriesberg, el control que ejercen las empresas sobre los datos y su acceso presenta desafíos de preservación, al limitar las actividades de productores, investigadores, curadores digitales y archivistas.[100]

Acker y Kriesberg dividen estos desafíos en tres niveles. El primero, de infraestructura, apunta al comportamiento de las API y las dificultades de mantener la estabilidad (*fixity*) de los contenidos, uno de los principios de la preservación digital y del modelo *Open Archival Information System* (OAIS). Esto porque las API generan flujos de actividades, es decir, la descripción de lo que realizan los usuarios en estos espacios.[101] Aunque en el acopio se genera un fichero en formato JSON, esto no asegura la preservación de objetos a los que un tuit hace referencia

98 Amelia Acker y Adam Kreisberg, "Social media data archives in an API-driven world", *Archival Science* 20, n.º 2 (2020).

99 Michelle Caswell, Marika Cifor y Mario H. Ramirez, "To Suddenly Discover Yourself Existing": Uncovering the Impact of Community Archives", *The American Archivist* 79, n.º 1 (2016); Christian Kelleher, "Archives Without Archives: (Re) Locating and (Re) Defining the Archive Through Post-Custodial Praxis", *Journal of Critical Library and Information Studies* 1, n.º 2 (2017).

100 Acker y Kreisberg, "Social media data", 108.

101 *Ibid*., 110.

como, por ejemplo, un video o una foto. El segundo nivel es el de estructura, donde el desafío principal consiste en el control que tienen las empresas sobre la manera en que estructuran los datos, lo cual trae implicaciones de interoperabilidad.[102] Y tercero, está el nivel de contenido, enfocado en los desafíos asociados al acceso y uso de datos generados por los usuarios, creando cuestionamientos relacionados con la privacidad, el consentimiento y el control de las plataformas de redes sociales sobre estos contenidos.[103]

El uso de datos de redes sociales para investigación no está exento de críticas y advertencias. En el caso de Twitter, Bruns y Stieglitz presentan importantes observaciones dirigidas a ser cautelosos con la idea de que sea representativo de la sociedad. Por una parte, ponen de manifiesto las limitaciones en el acceso de datos de Twitter a través de API si se realiza de manera gratuita, debido a que esta red establece límites en cuanto al máximo de tuits por minuto que se pueden recopilar. Por otra parte, explican la importancia de comprender "el papel de *Twitter* en el ecosistema de los medios de comunicación antes de evaluar qué aspectos del debate público pueden representar y cuán bien lo puede hacer".[104] Por su parte, Treem *et al*. advierte que es necesario pensar si toda red social es social, esto debido a que un gran porcentaje de los usuarios de redes sociales son una "mayoría silente" que no participa activamente en las discusiones, sino que asumen un rol de observadores.[105]

El corpus de investigaciones que se han estado desarrollando sobre las redes sociales desde disciplinas como las Ciencias Sociales, los estudios de medios y las Ciencias de la Información formulan un reconocimiento sobre los impactos sociales de las plataformas de redes sociales, su inserción en debates públicos y en movimientos sociales, así como los peligros ante la presencia significativa de la desinformación y las prácticas del capitalismo de vigilancia que buscan predecir e incidir en

102 *Ibid*., 111.

103 *Ibid*., 113.

104 Axel Bruns y Stefan Stieglitz, "Twitter data: what do they represent?", *It-Information Technology* 56, n.° 5 (2014): 243.

105 Jeffrey W. Treem *et al*., "What We Are Talking About When We Talk About Social Media: A Framework for Study", *Sociology Compass* 10, n.° 9 (2016): 772.

el comportamiento de las personas. Este reconocimiento ha llevado a su vez al desarrollo de proyectos de preservación de contenidos de las redes sociales, insertándose en las prácticas de Archivado Web presentes desde finales del siglo xx.

Uno de estos proyectos que fueron anunciados con gran expectativa, pero que a su vez ha generado lecciones sobre las complejidades del archivado de redes sociales, es el acuerdo entre Twitter y la Biblioteca del Congreso de los Estados Unidos para que este último preserve todo el archivo de tuits públicos. El acuerdo anunciado el 14 de abril de 2010 estableció que Twitter proveerá a la Biblioteca del Congreso todos los tuits públicos desde el inicio de la red social (2006) a la fecha del acuerdo y que la red social continuará proveyendo tuits de manera periódica. Además de que solo se iban a preservar los tuits públicos, ambas partes acordaron condiciones sobre esta adquisición, entre las que se encuentra un embargo de seis meses para que la Biblioteca del Congreso provea acceso a los tuits, y que dicho acceso será permitido a investigadores "*bona fide*".[106] Este ambicioso proyecto de preservación digital se enfrentó a una serie de obstáculos que llevaron a la Biblioteca del Congreso a anunciar el 26 de diciembre de 2017 que a partir del 1 de enero de 2018 se detenía la adquisición de todos los tuits públicos, modificando la estrategia a una adquisición selectiva de tuits, siguiendo las políticas y prácticas de la institución sobre Archivado Web.[107]

Michael Zimmer divide estos obstáculos en dos categorías, una enfocada en las prácticas de archivado de Twitter y la otra en el desarrollo e implementación de políticas.[108] Sobre las prácticas, Zimmer apunta a la inmensa cantidad de tuits que se pretendía preservar. La primera adquisición de tuits del 2006 al 2010 consistió en 21 billones de tuits,

106 Michael Zimmer, "The Twitter Archive at the Library of Congress: Challenges for information practice and information policy", *First Monday* 20, n.° 7 (6 de julio de 2015), https://doi.org/10.5210/fm.v20i7.5619.

107 Gayle Osterberg, "Update on the Twitter Archive at the Library of Congress", *Library of Congress Blogs*, 26 de diciembre de 2017, https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/.

108 Zimmer, "The Twitter Archive".

y en diciembre de 2012 recibió 150 billones de tuits adicionales. Además, cada tuit contaba con 50 campos de metadatos. En cuanto a la curaduría y acceso de este archivo, al 2015 aún no quedaba claro cómo la Biblioteca del Congreso planteaba gestionar los metadatos y el medio para proveer acceso a los usuarios. Esto también presentó desafíos desde el punto de vista de las políticas. Por una parte, se encuentran las restricciones de acceso y uso estipulados en el acuerdo entre Twitter y la Biblioteca del Congreso. Por otra, las de cómo balancear el acceso a la información de los tuits y respetar la privacidad de los usuarios, quienes no dieron el consentimiento para que sus tuits sean preservados. Zimmer explica que un vocero de la Biblioteca del Congreso presentó el argumento de que los tuits recibidos son públicos. Sin embargo, considera también que este argumento "presume una falsa dicotomía que la información es estrictamente pública o privada, ignorando cualquier norma contextual que haya guiado la publicación inicial de la información en Twitter o cómo una persona comprenda el flujo de información de un tuit".[109] Axel Bruns y Katrin Weller apuntan a otro desafío que podemos relacionar con los hallazgos de Acker y Kriesberg sobre la contextualización de los contenidos de las redes sociales. Bruns y Weller explican que por la manera en que Twitter provee los tuits a la Biblioteca del Congreso, principalmente enfocándose en el texto, se pierde contenido importante como aspectos visuales y los contenidos que provienen de las URL que se incluyen en los tuits.[110]

Estos son desafíos a los que todo proyecto de archivado de redes sociales se enfrenta, independientemente del alcance y volumen de la información que se desee preservar. A lo largo de las siguientes páginas del libro profundizamos sobre estos y otros retos, así como sobre estrategias implementadas en proyectos de Archivado Web y de redes sociales, incluyendo el proyecto de preservación de contenidos digitales sobre el COVID-19 en México.

109 *Ibid*., párr. 33.

110 Axel Bruns y Katrin Weller, "Twitter as a first draft of the present: and the challenges of preserving it for the future", en *WebSci '16: Proceedings of the 8th ACM Conference on Web Science* (NY, USA: Association for Computing Machinery, 2016), 186, https://doi.org/10.1145/2908131.2908174.

III. Colecciones como datos

3.1. Unicidad y redundancia documental

Las prácticas documentales se fundamentan en la unicidad y la redundancia a partir del número de ejemplares de un material. En tanto que los archivos y los museos conservan documentos únicos e inéditos, las bibliotecas pueden resguardar varios ejemplares de un mismo libro. La copia como método de salvaguarda documental fue durante siglos un recurso para garantizar que un mayor número de personas tuvieran acceso a los libros. Así, durante el Imperio bizantino en el siglo IX, se produjeron copias y se posibilitó que estas publicaciones circularan y fueran leídas, mientras que el original se conservaba en la biblioteca. A partir del siglo XVII fue posible conservar uno o varios ejemplares de un solo libro.[111]

La redundancia es la abundancia o repetición de alguna cosa. En el caso de las publicaciones impresas se utiliza para referirse a las copias de un mismo documento. En el entorno digital se refiere a medidas de seguridad del almacenamiento digital. Esta cualidad incide en la posibilidad de que un documento pueda ser preservado a través del tiempo. En la medida que se conserven más copias de un material, distribuidas en diferentes lugares, es posible que la información permanezca y sea accesible a un mayor número de personas. Y si se aplican técnicas de redundancia a colecciones digitales se garantiza su permanencia aún en

111 Masanès, "Web Archiving".

condiciones de posibles desastres naturales, fallas tecnológicas o errores humanos.

Contrario a esta posibilidad, la unicidad se refiere a que un documento es inédito y no ha sido publicado en serie y sólo se conserva un ejemplar; en consecuencia, si este desaparece o se borra la información que contiene se vuelve irrecuperable.

Se estima que se conserva una de cada 40 obras de la antigüedad. Este promedio se incrementó en el siglo XVII a una de cada dos. En la actualidad, en la mayoría de los países la conservación de las publicaciones impresas es alta porque se valora al libro como un artefacto cultural que forma parte del patrimonio de los pueblos.[112]

La conservación de los materiales librarios es diferente a las alternativas que tienen las publicaciones de la Web. Este tipo de documentos son vulnerables porque son únicos y efímeros; aun cuando su naturaleza digital conlleva la posibilidad de que se genere un número infinito de copias de un mismo documento. En muchas ocasiones, su preservación depende del productor.

Los recursos publicados en la Web tienen un origen (el servidor) e identificador único. Es decir, su existencia depende de su fuente única. Lo que supone una diferencia con la impresión porque el libro una vez publicado permanece por sí mismo. En tanto que los servidores web pueden adaptar y cambiar de forma sistemática el contenido en un mismo URL.

Por ello, para preservar las páginas web y los contenidos publicados en redes sociales, la información acopiada debe adquirir una estructura o formato a través de la cual se codifica el objeto digital para que pueda ser procesado, conservado y sobre todo para que pueda estar disponible para su acceso a largo plazo. Cuando se realiza el Archivado Web y de redes sociales se han de tomar en cuenta los formatos estandarizados que garanticen la recuperación de la información, su interoperabilidad y conservación sustentable. Los contenidos de la Web y de redes sociales acopiados forman colecciones digitales o conjuntos de datos.

112 *Idem.*

3.2. El Archivado Web y la noción de colecciones como datos

El 27 de abril de 2018 la periodista Cristina Fallarás publicó un tuit proponiendo la etiqueta #cuéntalo para denunciar por Twitter casos de agresión sexual que mujeres han vivido.[113] La iniciativa de Fallarás surgió a raíz de la sentencia en el caso "La Manada", donde un grupo de cinco hombres violó a una joven de 18 años, y quienes fueron sentenciados a nueve años por abuso sexual y no por violación, causando protestas por lo leniente de tal sentencia.[114] Por medio del movimiento *#Cuéntalo* 2.8 millones de mujeres compartieron sus testimonios sobre violencia sexual. Del 27 de abril al 13 de mayo de 2018, Vicenç Ruiz Gómez y Aniol Maria Vallès, miembros de la Asociación de Archiveros de Cataluña, realizaron el acopio de dos millones de tuits con la etiqueta *#Cuéntalo*.[115] Este acopio redundó en un proyecto realizado en colaboración con las periodistas Fallarás y Karma Peiró, y con el Centro Nacional de Supercomputación en Barcelona, el cual consistió en el análisis y visualización de 160 000 tuits originales del set de datos.[116] Al entrar al sitio http://www.proyectocuentalo.org se puede navegar por una visualización en forma de círculo que organiza los tuits en tres tipos de mensajes: testimonio, mensajes de apoyo y otros. El círculo representa un reloj de 24 horas, organizando los tuits por la hora en que fueron publicados.[117]

Este tipo de análisis y visualización ejemplifica la noción de colección o archivo como datos. Esta perspectiva se basa en la integración de

113 efe, "Sentir miedo al volver a casa por las noches…", *Heraldo*, 29 de abril de 2018, https://www.heraldo.es/noticias/nacional/2018/04/28/miles-mujeres-relatan-sus-agresiones-sexuales-twitter-animadas-bajo-lema-cuentalo-1240464-305.html.

114 "La Manada: ¿cómo acabó el repudiable caso de violación grupal que conmocionó a España?", *El Comercio*, 20 de octubre de 2020, https://elcomercio.pe/mundo/europa/la-manada-como-acabo-el-repudiable-caso-de-violacion-grupal-que-conmociono-a-espana-noticia/. El Tribunal Supremo retomó el caso y el 21 de junio de 2019 concluyó que el grupo cometió violación, elevando la sentencia a 15 años de cárcel.

115 Vicenç Ruiz Gómez y Aniol Maria Vallès, "#Cuéntalo: the path between archival activism and the social archive(s)", *Archives & Manuscripts* 48, n.º 3 (2020): 281.

116 *Idem*.

117 Cristina Fallarás *et al.*, El Proyecto #Cuéntalo, https://www.bsc.es/viz/cuentalo/.

métodos computacionales para el análisis y la representación de archivos o colecciones digitales.[118] Estos archivos o colecciones no se consideran datos solo por el hecho de que sean digitales, sino que requieren de procesos de curaduría digital para permitir su análisis y visualización por métodos computacionales.[119] En el contexto del Archivado Web la noción de colecciones como datos se identifica en los procesos de acopio y curaduría, y de la comprensión de los contenidos preservados como datos estructurados que permiten la reproducción de la información, así como su análisis desde diversos métodos. En este capítulo profundizamos en relación con los datos estructurados, explicando a fondo el formato WARC de preservación digital de contenidos web y el JSON, utilizado para el acopio de tuits por medio de las API.

Es importante reflexionar sobre las implicaciones de esta perspectiva en las aproximaciones teóricas y prácticas de las ciencias de la información. Devon Mordell, hablando desde la perspectiva archivística, alerta sobre los peligros de pensar que la implementación de métodos computacionales para el manejo de archivos digitales nos llevaría a la neutralidad y la objetividad: "ni el archivista ni el algoritmo existen sin sus propias historias, suposiciones o sesgos conformados por relaciones desiguales de poder".[120] Por su parte, el historiador y especialista en humanidades digitales, Nicolás Quiroga, aboga por la importancia de estudiar y comprender cómo se desarrollan las herramientas digitales para la aplicación de métodos computacionales en la curaduría y análisis de archivos. Tomando como ejemplo el caso del archivo web de Geocities, que cuenta con más de 15 millones de documentos con extensión html, Quiroga reflexiona sobre el uso de métodos computacionales y el desarrollo de habilidades sobre estructuras de datos y codificación:

118 Cory Lampert y Emily Lapworth, "What do we mean by «collections as data» (CAD)?", UNLV University Libraries, enviado el 2 de marzo de 2020, https://www.library.unlv.edu/whats-new-special-collections/2020/2020-03/what-do-we-mean-collections-data-cad-cory-lampert-emily#:~:text="Collections%20as%20data"%20is%20the,or%20people%20that%20are%20named.

119 Devon Mordell, "Critical Questions for Archives as (Big) Data", *Archivaria* 87 (2019): 144.

120 *Ibid*., 149.

> (…) frente a colecciones "ilegibles", la encrucijada interpretativa no está dada por los problemas de las máquinas para comprender la experiencia humana, sino por el problema de las personas para comprender los procesos computacionales. En los últimos años eso ha impulsado la conversación sobre metodologías digitales y también el "cacharreo", el uso de esas herramientas. Pero no ha sido igual de notable el interés por el análisis de los algoritmos que usan esos programas. Sin ese análisis, estamos cerca de ser usuarios de instrumentos desconocidos para analizar repositorios desconocidos.[121]

3.3. La Web como documento: formato WARC

La Web es la principal aplicación para publicar información digital en Internet. Día a día se producen y actualizan páginas web en una amplia variedad de géneros como son los blogs, las *wikis*, redes sociales como Twitter, Facebook, Linkedln, Tik Tok, entre otros. Estos son objetos digitales hipermedia, definidos así por la diversidad de lenguajes que emplean. Son materiales que no están aislados, se encuentran interconectados; derivado de lo cual, su extensión es compleja de determinar porque a partir de una página se establecen enlaces (*links*) entre los contenidos. Son dinámicos y efímeros porque los datos pueden ser editados y modificados constantemente.[122]

Para salvaguardar la información publicada en una página web es necesario poner en marcha procesos de Archivado Web, es decir de rastreo o recolección de datos; términos que en inglés corresponden a *Web Archiving*, *Webcrawling* o *harvesting*.[123] Los datos recopilados se estructuran y almacenan en un formato a través del cual puedan ser desplegados y consultados sin que para ello sea necesario utilizar un archivo complementario.

121 Nicolás Quiroga, "Interpretación histórica y objetos digitales: consideraciones a partir de ejemplos concretos", *Vegueta: Anuario de la Facultad de Geografía e Historia* 22, n.º 1 (2022): 50.

122 Masanès, "Web Archiving".

123 Arnoud Goos, "Archiving broadcasters' websites a discussion of web archiving as context to the radio and television collection", en *2015 Digital Heritage, Granada, Spain,* ieee *Xplore* (2015): 627-630.

Las redes sociales forman parte de la Web. Por lo tanto, en estricto sentido el Archivado Web comprende a las redes sociales. Sin embargo, en esta publicación se tratan de manera separada para ahondar en sus especificidades.

En 1996, Internet Archive comenzó a utilizar el formato ARC (ARC_IA) para almacenar secuencias de bloques de contenidos recopilados de la Web.[124] Esta forma de codificación de los datos publicados en la Web se generó como producto del empleo de Heritrix, primer *software* rastreador de páginas web.

En 2009, se estableció el formato WARC (Web ARChive), a través del cual se especifica un método para combinar múltiples recursos digitales e información agregada en un archivo. El WARC es una versión actualizada del formato ARC_IA.

El formato WARC fue adoptado como norma ISO 28500:2009 para el archivado web[125] para conservar y preservar documentos recopilados en Internet.[126] Este formato crea un paquete de datos con diferentes tipos de contenidos y mantiene la relación entre las páginas web acopiadas. Fue utilizado primero por Internet Archive y después por los miembros del International Internet Preservation Consortium (IIPC). En la actualidad es el formato más empleado para la preservación de páginas web y redes sociales.

Los WARC son formatos de objetos digitales complejos porque pueden ser contenedores de otros WARC. Es decir, concatena archivos en varios formatos, entre otros se pueden citar PDF, MP3, MXF y además, pueden almacenar información que esté en contenedores como ZIP, GZIP, TAR o RAR. Son producidos por rastreadores, *proxies* y otras utilidades que recuperan archivos de un servidor web activo y pueden contener el HTML,

124 "Sustainability of Digital Formats: Planning for Library Congress Collection", Library of Congress, consultado el 5 de junio de 2023, https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml.

125 Kristine Hanna, "El Modelo de Ciclo de Vida del Archivado Web", en *Anuario AC/E de cultura digital*, editado por AC/E Acción Cultural Española (2014), https://www.accioncultural.es/media/Default%20Files/activ/2014/multimedia/anuario%20ace/Anuario_ACE_cultura_digital_2014.pdf.

126 Clément Oury, "WARC implementation guidelines. Contribution from WARC usage task force", IIPC, publicado el 27 de enero de 2009, https://netpreserve.org/resources/WARC_Guidelines_v1.pdf.

JS, CSS y otros elementos estructurales que los navegadores web necesitan leer para representar el contenido del sitio a los usuarios. Los WARC contienen metadatos técnicos y de origen de procedencia de modo que los sitios pueden leerse y representarse en experiencias de navegación web en vivo tal y como eran en el momento de su recopilación.[127]

Los archivos WARC suelen contener vastos volúmenes de datos de la Web, cuyo manejo puede ser complejo derivado del peso por la magnitud del archivo. Por lo cual, en 2021, Webrecorder creó el formato WACZ (Web Archive Collection Zipped),[128] para el empaquetamiento de datos y metadatos a fin de que las tareas de acopio y almacenamiento de archivos web sean más fáciles y rápidas de llevar a cabo.

De acuerdo con Ilya Kreymer y Emma Dickson, creadores de Webrecorder, el formato WACZ pretende conectar y comunicar las colecciones de archivos web, contar con información de contexto –describiendo cuándo y cómo se creó el archivo–, para interpretar, interactuar y posibilitar que la información se cargue de forma dinámica desde un *host* remoto sin necesidad de descargar el archivo completo.[129]

Los archivos WACZ son archivos ZIP que contienen los archivos WARC sin procesar. Este tipo de archivo puede ser leído bajo demanda, a través de la red, sin necesidad de descargar el archivo completo. Los WACZ incluyen todo lo necesario para crear y alojar una colección de archivos web: interfaz de acceso aleatorio a todos los datos en bruto, lista de páginas de entrada al archivo y metadatos editables definidos por el usuario sobre la colección de archivos web. Así como los datos de texto completo extraídos de las páginas web, que pueden ser introducidos en motores de búsqueda como Solr o cargados sobre la marcha junto con la reproducción. En mayo de 2023, el formato WACZ se actualizó.[130]

127 Internet Archive, "The stack: An introduction to the WARC file", https://ait.blog.archive.org/post/the-stack-warc-file/.

128 "Web Archive Collection Zipped", Library of Congress, consultado el 5 de junio de 2023, https://www.loc.gov/preservation/digital/formats/fdd/fdd000586.shtml.

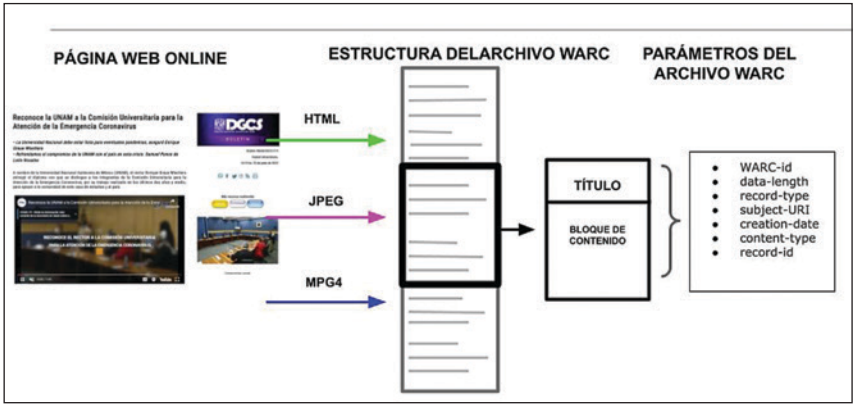
129 Ilya Kreymer y Emma Dickson, "Announcing WACZ Format 1.0", Webrecorder Web archiving for all!, publicado el 18 de enero de 2021, https://webrecorder.net/2021/01/18/wacz-format-1-0.html

130 "Sustainability of Digital", Library of Congress.

A continuación, se ofrece la imagen de una página web de la UNAM que fue seleccionada y acopiada por el doctor Jonathan Hernández Pérez. La página corresponde al Boletín de la DGCS-UNAM, en donde el rector Enrique Graue Wiechers entregó el diploma con que se distingue a los integrantes de la Comisión Universitaria para la Atención de la Emergencia Coronavirus. La publicación acopiada incluye en el archivo hipertexto, gráficos, fotografías y un video.

Fuente: Universidad Nacional Autónoma de México.

En el siguiente recuadro se presenta un esquema que muestra cómo se capturan las diversas informaciones de la página en formato WARC.



Fuente: Elaboración propia con información de la UNAM.

3.4 Conjunto de datos de redes sociales (Twitter)

El manejo de datos en redes sociales como Twitter funciona dentro del ecosistema conocido como la interfaz de programación de aplicaciones o Application Programming Interface (API). La API consiste en una serie de protocolos utilizados para comunicar *softwares* entre sí con el propósito de consultar datos, analizar respuestas y enviar instrucciones.[131] Las redes sociales funcionan a través de los API para compartir datos y como parte de su modelo de negocios. Un ejemplo que se utiliza comúnmente es cuando como usuarios decidimos acceder a un sitio web utilizando nuestra cuenta de Facebook o de Gmail. Al acceder con una de estas cuentas estamos enviando una instrucción de una plataforma (Facebook o de Gmail) a otra plataforma para lograr el acceso.

En el contexto de Twitter, este ecosistema impacta la manera en que se realiza la captura para su preservación, así como su formato de almacenamiento. Por una parte, es posible almacenar tuits en el formato WARC o WARCZ si la captura se realiza a través del método de rastreo de datos (*web crawling* o *harvesting*). Esta estrategia permite preservar el estilo visual y comportamiento de los tuits ("*look and feel*"). En este caso, el propósito principal es almacenar y recrear la funcionalidad de la página al momento de la captura.[132] Sin embargo, se pueden perder los datos estructurados que conforman los tuits y que permiten la realización de visualizaciones y análisis de *big data*.

Esta es una de las razones por las que la estrategia más empleada es el acopio de datos estructurados, el cual consiste en realizar el acopio por medio de la API. En este caso, el propósito principal es preservar la información derivada de los datos sin procesar ("*raw data*") que contiene cada tuit.[133] A través de esta estrategia se pueden acopiar grandes cantidades de tuits, presentando mayores oportunidades para realizar

131 "Interfaz de programación de aplicaciones (API)", Invatati Afaceri, publicado el 18 de noviembre de 2022, https://invatatiafaceri.ro/es/diccionario-financiero/interfaz-de-programacion-de-aplicaciones-api/.

132 Zefi Kavvadia, "An Overview of Social Media Archiving Tools" (versión 1.0 diciembre 2020),15, Zenodo, publicado el 2 de febrero de 2021, https://doi.org/10.5281/zenodo.4493594.

133 *Idem*.

análisis de datos y visualizaciones aplicando métodos computacionales.[134] Para Twitter, los tuits acopiados por este método son almacenados en el formato JSON.

JSON significa JavaScript Object Notation y es un formato para intercambio de datos que por ser en texto facilita compartir los datos entre computadoras.[135] En el caso de Twitter, el archivo JSON de tuits abarca el contenido de los tuits, así como otros metadatos como el identificador de cada tuit, fecha de creación, identificador de usuario y acciones realizadas como retuits.[136] El siguiente ejemplo muestra la estructura JSON de un tuit.

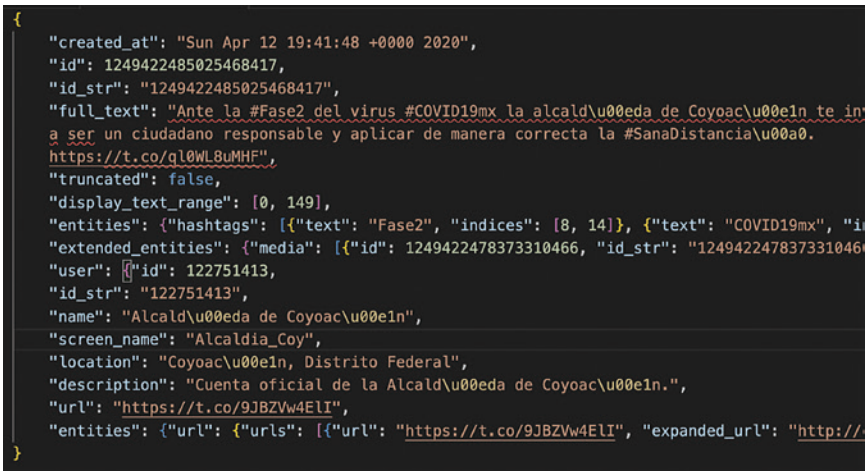
La primera imagen muestra un tuit publicado por la Alcaldía de Coyoacán el 5 de abril de 2020:

134 *Ibid*., 16.

135 "JSON – Introduction", W3 Schools, consultado el 25 de enero de 2023, https://www.w3schools.com/js/js_json_intro.asp.

136 Sara Day Thomson, *Preserving Social Media* (Great Britain: Digital Preservation Coalition, 2016), 16.

La segunda imagen muestra los datos estructurados del tuit en formato JSON, donde podemos ver que la sintaxis JSON consiste en el nombre dado a un campo de información (ejemplo "*created_at*") y el valor asignado a ese campo ("Sun Apr 12 19:41:48 +0000 2020").

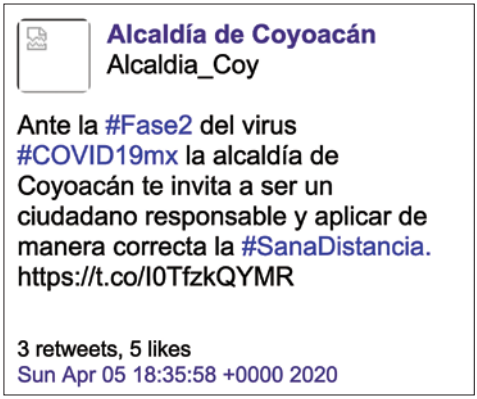


Este ejemplo representa solo un tuit. Cuando se realiza el acopio de tuits a través de la API, el archivo JSON suele incluir miles de tuits. El ejemplo del tuit de la Alcaldía de Coyoacán forma parte de un archivo JSON que cuenta con más de 250 000 tuits.

Desde el 2014, la Biblioteca del Congreso de los Estados Unidos publica un informe anual sobre formatos de preservación digital para diversos tipos de documentos, uno de los cuales son archivos web. El informe 2022-2023 indica que el formato de preservación preferente es WARC. El formato WARCZ es considerado como aceptable para preservación digital. Sin embargo, no se presentan recomendaciones particulares de preservación digital para redes sociales. Al respecto, la Biblioteca del Congreso reconoce que contenidos en la Web, incluyendo los de las redes sociales, utilizan formatos de reproducción que son diferentes al formato de preservación digital recomendado. Y añade: "dado que el enfoque de este documento es la preservación y acceso a largo plazo,

los formatos recomendados dan prioridad a este enfoque y se incluyen recomendaciones de mejores prácticas para permitir la preservación de contenidos web".[137]

En el caso de Twitter, este escenario se ve reflejado por medio de los enlaces que se incluyen en los tuits. En otras palabras, si un tuit acopiado contiene un video, foto o un enlace a otra fuente de información, el archivo JSON captura la URL, pero no el contenido de esa URL. Siguiendo con el ejemplo anterior de la Alcaldía de Coyoacán, el tuit incluye una imagen, la cual no es reproducida cuando se realiza una conversión de JSON a HTML.



Comunidades de práctica de archivado de redes sociales han desarrollado *scripts* con el lenguaje Python para extraer estos contenidos y realizar un ejercicio de archivado web. Por ejemplo, como parte de la librería de líneas de comando de *twarc*, se desarrolló un *script* que extrae las URL del archivo JSON y genera una versión WARC del contenido de esa URL.[138]

137 "Library of Congress Recommended Formats Statement 2022-2023", Library of Congress, consultado el 16 de junio de 2023, https://www.loc.gov/preservation/resources/rfs/RFS%202022-2023-ArchivalOnly.pdf; "X. Web Archive", Library of Congress, consultado el 16 de junio de 2023, https://www.loc.gov/preservation/resources/rfs/webarchives.html.

138 "DocNow/twarc", GitHub, consultado el 23 de marzo de 2023, https://github.com/DocNow/twarc/blob/main/utils/media2warc.py.

IV. Curaduría y selección de contenidos

4.1 El ciclo de vida en el Archivado Web y de redes sociales

El Archivado Web, de acuerdo con el International Internet Preservation Consortium (IIPC), es el proceso mediante el cual se recolectan porciones de la World Wide Web para crear colecciones en formatos de archivo a fin de que puedan ser consultados y usados.[139] En esta definición se enuncia la recolecta o acopio de porciones de la Web, primera tarea intelectual que supone la necesidad de seleccionar fragmentos de la Web. Después, se advierte la necesidad de que los materiales acopiados sean conservados en formatos de archivos (ARC, WARC o WARCZ); para que puedan ponerse a disposición de los usuarios para su consulta y uso de forma sustentable y a largo plazo. Así se establecen de forma simplificada los procesos y flujos que intervienen en el *Ciclo de Vida del Archivado Web*.

Hace más de una década el grupo de trabajo de Archive-It, un servicio de archivado web administrado por el Internet Archive, formuló el modelo de *Ciclo de Vida del Archivado Web*.[140] La elaboración del modelo surgió a partir de los requerimientos de instituciones que realizaban prácticas de archivado web y utilizaban Archive-It, y su objetivo fue atender estas necesidades y concientizar sobre la importancia del

139 "About Archiving", International Internet Preservation Consortium, consultado el 5 de junio de 2023, https://netpreserve.org/web-archiving/about-archiving/.

140 Hanna, "El Modelo".

Archivado Web como parte integral de la preservación digital.[141] En este sentido, el modelo busca proveer un documento fundacional que sirva como guía para el desarrollo de programas de archivado web.[142]

El modelo de *Ciclo de Vida del Archivado Web* expresa la relación y naturaleza reiterativa entre los procesos y flujos cotidianos de trabajo que intervienen en la creación de las colecciones web que ocupan el centro del modelo. Rodean la creación de colecciones de la Web los procesos de valoración y selección, definición del alcance, captura de datos, almacenamiento y organización, seguridad y cuidado de la calidad de los contenidos. El siguiente nivel lo ocupan los metadatos e información descriptiva que circundan todo el modelo porque son recopilados en múltiples etapas y con diversos propósitos. A continuación, se sitúan las decisiones de alto nivel que guían todos las etapas y aspectos involucrados en el Archivado Web: conservación, recursos y flujos de trabajo, gestión de riesgos, acceso, utilización y reutilización; así como la visión y objetivos para definir el por qué y los alcances del Archivado Web.[143] El último círculo corresponde a las políticas determinadas en función de las etapas y los flujos de trabajo, metadatos y decisiones de alto nivel que intervienen en el Archivado Web.

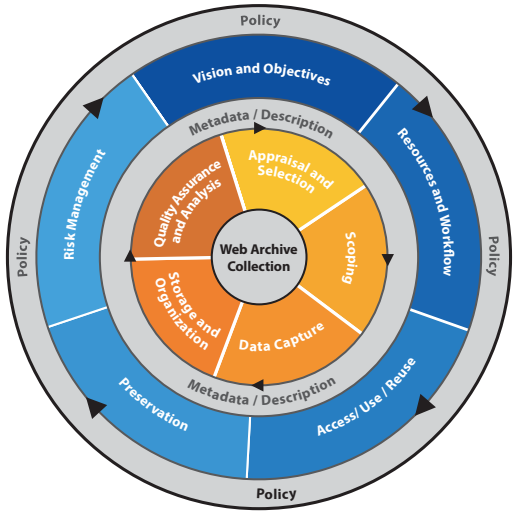
La lectura del modelo puede iniciar desde la parte externa a partir de las políticas o bien en función de la creación de las colecciones web. En el primer caso se observa cómo el marco general de las políticas conlleva a la creación de colecciones. Y en el segundo, se establece una relación de lo específico, creación de colecciones al marco general determinado por la definición de políticas. De forma independiente al modo en que pueda ser interpretado este modelo, ofrece una idea clara de los procesos intelectuales y tecnológicos que intervienen en el Archivado Web. La instauración de este modelo fue resultado de una investigación en torno a la forma en que se llevan a cabo los procesos de Archivado Web en diversas instituciones.

141 Molly Bragg y Kristine Hanna, "The web archiving life cycle model", The Archive-It Team Internet Archive, publicado en marzo de 2013, https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf y https://archive.org/details/WALCM.

142 Jefferson Bailey, director de Servicios de Datos de Internet Archive, entrevistado por Joel Antonio Blanco-Rivera, el 14 de julio de 2023.

143 Hanna, "El Modelo".

Modelo Ciclo de Vida del Archivado Web



Fuente: Internet Archive.

¿Cómo se ha implementado el modelo del *Ciclo de Vida del Archivado Web*? En el 2017 el National Digital Stewardship Alliance (NDSA) realizó su cuarta encuesta sobre prácticas de archivado web en Estados Unidos.[144] La encuesta incluyó preguntas relacionadas con el modelo desarrollado por Internet Archive. Se analizaron respuestas de 119 instituciones, de las cuales el 61 % provienen de instituciones de educación superior. Dos preguntas de la encuesta mencionan directamente el modelo de *Ciclo de Vida del Archivado Web*, indagando sobre las dimensiones donde se ha logrado mayor y menor progreso. Respecto a las dimensiones donde alcanzaron mayor progreso, las tres que recibieron el más alto porcentaje fueron Captura, Valoración y Selección, y Alcance (*Scoping*);[145] mientras que las tres dimensiones con menor progreso

144 Las cuatro encuestas pueden ser consultadas en "National Digital Stewardship Alliance (NDSA) / Web Archiving Survey", OSF, última actualización el 6 de enero de 2022, https://osf.io/4ytb2/. En el 2022 el grupo de trabajo realizó su quinta encuesta, la cual no ha sido publicada.

145 Matthew Farrell *et al*., *Web Archiving in the United States: A 2017 Survey*, An NDSA Report (National Digital Stewardship Alliance, octubre 2018), 12.

fueron Acceso, Uso y Reuso, Metadatos y Descripción, y Control de Calidad y Análisis.[146]

En relación con las dimensiones de menor progreso fueron los temas de acceso, uso y metadatos. El estudio indica que existe falta de claridad sobre cómo son utilizados los materiales, además de que las instituciones han encontrado dificultades en la implementación de prácticas de descripción bibliotecológica y archivística a los archivos web.[147] Estas observaciones coinciden con otros estudios realizados en esa época que identifican desafíos similares. El estudio de Gail Truman para la Biblioteca de la Universidad de Harvard identificó retos por parte de las instituciones y de usuarios en temas de acceso y metadatos.[148] Los usuarios entrevistados para este estudio señalaron la necesidad de desarrollar habilidades de análisis de datos utilizando métodos computacionales,[149] así como contar con más información por parte de las instituciones sobre la conformación de los archivos web, brindando datos sobre las políticas que inciden en la selección, acopio, descripción y acceso a los contenidos web.[150]

El grupo de trabajo de Archive-It también ha realizado encuestas con instituciones que utilizan el servicio. Estas encuestas, llamadas S*tate of the* WARC *Report*, se enfocan en aspectos particulares del modelo, especialmente relacionados con prácticas de almacenamiento y metadatos.[151] Por ejemplo, en un informe publicado en 2020 se observó que más de la mitad de las instituciones encuestadas no almacenan una copia del archivo web conformado en Archive-It, dependen solo del almacenamiento que proporciona este servicio.[152] Jefferson Bailey, director de

146 *Ibid*., 13.

147 *Idem*.

148 Gail Truman, *Web Archiving Environmental Scan*, Harvard Library Report (enero de 2016), https://dash.harvard.edu/handle/1/25658314.

149 *Ibid*., 30.

150 *Ibid*., 31.

151 Bailey, entrevistado por Blanco-Rivera.

152 "State of the WARC Report: Web archive management and preservation in 2019-20", *Archive-It*, 26 de mayo de 2020, https://ait.blog.archive.org/post/state-of-the-warc-2020/ y https://archive-it.org/blog/files/2020/05/State-of-the-WARC_2020.pdf.

Servicios para Archivado y Datos (Archiving and Data Services), área que está a cargo de la encuesta, explica que una de las razones por las cuales la mayoría de las instituciones encuestadas no cuentan con una copia local de almacenamiento es que no ven viable almacenar y gestionar grandes cantidades de datos por las limitaciones de recursos.[153] Estas limitaciones incluyen el no contar con personal enfocado en el programa de archivado web y el que no se vea este como parte del programa institucional de desarrollo de colecciones y preservación digital.[154]

Bailey entiende que, a más de diez años de su publicación, el modelo *Ciclo de Vida del Archivado Web* necesita una revisión para actualizarlo al estado actual. Entre los temas principales están las complejidades vinculadas a la preservación de contenidos de redes sociales y el impacto de las API en los flujos de información. Bailey presenta, como ejemplo, el acopio de tuits y la toma de decisiones sobre selección, acceso y uso de los mismos, explicando que a diferencia de los modelos tradicionales de adquisición de fondos documentales en los proyectos de Archivado Web y de redes sociales es muy probable que no se establezca comunicación con los creadores de contenido. "La gente usa las redes sociales y no piensan en la posibilidad de que sus contenidos sean acopiados y preservados", señala Bailey, lo cual tiene implicaciones sobre acceso y uso:

> Pienso que esto no está muy elaborado en el modelo de ciclo de vida original. El modelo habla de permisos, ¿vas a preguntar a las personas si puedes archivar su sitio web, vas a notificar, pero no a solicitar autorización, vas a hacer el acopio y luego restringir el acceso o eliminarlo? Creo que esos temas son más complejos [actualmente] por las redes sociales.[155]

Otra razón es cómo ha evolucionado la manera en que los usuarios desean interactuar con los archivos web. "Hay más personas interesadas

153 Bailey, entrevistado por Blanco-Rivera.
154 *Idem*.
155 *Idem*.

en utilizar los archivos web más allá de consultarlos en el reproductor de sitios web archivados", explica Bailey.[156] Esto incluye el uso de prácticas de minería de datos y otros métodos computarizados vinculados a la idea de colecciones como datos. Finalmente, otro aspecto que no se explica más a fondo en el modelo es el de la colaboración entre instituciones y creadores de contenido o grupos comunitarios para desarrollar proyectos de Archivado Web.[157]

Sobre este elemento de colaboraciones, Bailey lo ve como una fortaleza en las prácticas actuales de Archivado Web. Cuando se publicó el modelo en el 2013, la mayoría de las instituciones involucradas eran bibliotecas académicas de universidades grandes, así como bibliotecas nacionales europeas. Actualmente, se han diversificado los perfiles de instituciones que realizan proyectos de Archivado Web, los cuales han traído además iniciativas colaborativas con otras instituciones y con organizaciones y creadores de contenido.

> La diversidad de las instituciones es radicalmente diferente de manera muy positiva en comparación a cómo era originalmente. Las universidades suelen tener enfoques muy específicos en el desarrollo de colecciones (…). Pero si vemos el trabajo de las *historical societies*, museos, bibliotecas públicas y archivos comunitarios e iniciativas de base, es abundante. Es un alcance más diverso.[158]

Con base en una revisión crítica del modelo *Ciclo de Vida del Archivado Web*, el análisis de bibliografía publicada en torno al tema y la sistematización del proceso de Archivado Web desarrollado en el marco del *Proyecto PAPIIT IT 400121 Preservación digital de contenidos publicados en portales web y redes sociales. Del acopio a la difusión de colecciones digitales sobre COVID-19 en México*, en este libro se sugieren y desarrollan cuatro procesos documentales para el Archivado Web y de redes so-

156 *Idem*.
157 *Idem*.
158 *Idem*.

ciales. Estos corresponden a flujos de trabajo y actividades intelectuales y físicas que intervienen en la preservación de este tipo de materiales. Fueron desarrollados para ser de utilidad a archivistas, bibliotecólogos, documentalistas, curadores y profesionales de la información interesados en iniciar programas de archivado web y de redes sociales.

Los procesos son:

- Curaduría o selección de contenidos
- Acopio
- Gestión y almacenamiento digital
- Acceso y reaprovechamiento documental

Cada uno es abordado por capítulo. En el primero se profundiza en torno al proceso de curaduría y selección, donde se explica la vinculación del archivado web con la curaduría digital, así como el proceso intelectual que deriva en el establecimiento de criterios de selección de contenidos web para la conformación de un archivo web.

4.2. Curaduría

El Archivado Web y de redes sociales es disruptivo con los métodos de trabajo documental empleados por archivistas, bibliotecarios y profesionales de la información en las tareas de salvaguarda documental. Las tareas repetitivas que privaron durante décadas para conservar en las mejores condiciones sendas colecciones analógicas se han modificado por la irrupción de documentos de origen digital, como son las páginas web y las redes sociales.

La naturaleza documental de este tipo de materiales requiere de la incorporación de procesos documentales distintos; además, conlleva la inserción de nuevos conceptos para nombrar procesos que antes no se desarrollaban pero que son necesarios en el manejo de este tipo de materiales. El uso de estos términos puede generar equívocos.

Un ejemplo de ello son las nociones de curaduría y preservación que se emplean de forma indistinta y confusa[159] para hacer referencia al ciclo de vida digital de los datos. Es decir, a las etapas a través de las cuales pasan los objetos digitales, desde su acopio hasta su acceso. Los términos se utilizan para nombrar procesos similares. La preservación digital se ocupa de garantizar la permanencia tanto de contenidos y metadatos que han sido digitalizados como de aquellos que son de origen digital. Mientras que la curaduría digital se emplea para referirse a la gestión activa para mantener los datos nativos digitales como son las páginas web.[160] En ambos casos, se pretende mantener la autenticidad, fiabilidad y accesibilidad de los objetos y datos digitales a largo plazo.

La curaduría digital se enuncia como "la actividad a través de la cual es posible mantener, preservar y agregar valor a los datos de investigación a lo largo de su ciclo de vida".[161] La noción formulada en el ámbito de la investigación científica pretende vincular al creador y al usuario de la información.[162] El propósito original de la curaduría digital es incrementar el valor de los datos y ponerlos a disposición de nuevas investigaciones.

En este sentido, una de las características principales de la curaduría digital es la implementación de procesos activos con los creadores de documentos con el fin de mantener su uso y re-uso.[163]

La vinculación del archivado web con la curaduría digital proviene además de la noción de archivo o colección como datos. Esta perspectiva se basa en la idea de que archivos o colecciones digitales pueden

159 Gillian Oliver y Ross Harvey, *Digital Curation*, 2.ª ed. (Chicago: ALA Neal-Schuman, 2016), 240.

160 Joel Blanco-Rivera, "Curaduría digital y la preservación de contenidos web: creando una colección de tuits sobre la huelga de la Universidad de Puerto Rico", Encuentro Latinoamericano de Bibliotecarios, Archivistas y Museólogos. Revalorizando el Patrimonio en la era digital, 9-13 de octubre de 2017, https://www.institutomora.edu.mx/EBAM/2017/Ponencias/Curaduria%20digital%20y%20la%20preservacion%20de%20contenidos%20web%20creando%20una%20coleccion%20de%20tuits%20sobre.pdf.

161 "What is digital curation?", Digital Curation Centre, consultado el 25 de mayo de 2023, https://www.dcc.ac.uk/about/digital-curation.

162 Oliver y Harvey, *Digital Curation*.

163 Elizabeth Yakel, "Digital Curation", *OCLC Systems & Services: International digital library perspectives* 23, n.º 4 (2007): 338.

ser utilizados como datos para el análisis por medio de métodos computacionales.[164] La investigación del historiador Ian Milligan sobre GeoCities[165] es un ejemplo de esta noción de colección como datos. GeoCiteis fue fundada en 1994 como uno de los primeros servicios de alojamiento web (*web hosting*).[166] Una de sus características más importantes era la organización de los sitios web por temas o "vecindarios", permitiendo al usuario seleccionar el "vecindario" donde se alojaría su sitio web. El servicio fue adquirido por Yahoo en 1999 y diez años después fue descontinuado.[167] Milligan aplicó métodos computacionales para el análisis de grandes cantidades de datos derivados del archivo web de GeoCities.

En relación con las prácticas de archivado web se puede utilizar el término curaduría digital para referirse al proceso intelectual inicial a través del cual se forman las colecciones digitales y se les confiere un valor de uso potencial, desde el momento mismo en que se conceptualiza su creación. Este proceso intelectual sirve de base para la implementación de actividades de selección, acopio, descripción y acceso al archivo web conformado, con el fin de permitir el uso y re-uso de los datos que conforman este archivo.

4. 3. Selección

La selección es una actividad cotidiana que se realiza en los archivos para priorizar tareas de conservación, catalogación, digitalización, entre otros procesos documentales. Ante el creciente volumen de contenidos, sobre todo de origen digital, la selección deviene en una tarea intelectual de relevancia para la preservación de contenidos digitales. En muchas instituciones de la memoria se cuenta con un Comité crea-

164 Lampert y Lapworth, "What do we mean by".

165 Milligan, "La historia en la era".

166 Gita Jackson, "The Geocities Archive is Bringing the Early Internet to Life", Vice, publicado el 27 de enero de 2020, https://www.vice.com/en/article/n7jzgm/the-geocities-archive-is-bringing-the-early-internet-to-life.

167 "GeoCities", AT Archiveteam, última modificación el 20 de enero de 2023, https://wiki.archiveteam.org/index.php/GeoCities#:~:text=GeoCities%20was%20a%20once%20very,on%20the%20World%20Wide%20Web.

do ex profeso para aplicar criterios de selección y con ello, encauzar de manera más eficiente los recursos humanos, técnicos y económicos destinados a la preservación.

En el Archivado Web y de redes sociales la selección determina las características y delimita el alcance de las colecciones de datos de origen digital. La selección sustituye a la adquisición, tarea destinada a la creación de colecciones documentales librarias, sonoras y audiovisuales. Las colecciones de páginas web y de redes sociales no se pueden comprar, donar o entregar en comodato a bibliotecas y archivos. Este tipo de colecciones deben crearse con base en principios curatoriales y criterios de selección.

La Unesco publicó en 2016 las *Directrices Unesco/PERSIST sobre selección del patrimonio digital para su conservación a largo plazo* para ayudar a las instituciones culturales en la salvaguarda de este tipo de herencia antes de que sea demasiado tarde.[168] En ese documento señaló tres estrategias para acopiar el patrimonio digital: la recopilación exhaustiva, el muestreo representativo y la selección.

La recopilación exhaustiva es el modo para acopiar la mayor cantidad de información sobre un tema, periodo o región geográfica determinada. El depósito legal es el ejemplo más significativo en esta modalidad. El muestreo representativo es un enfoque que se puede poner en marcha cuando una institución no cuenta con todos los recursos necesarios para preservar todo. Esta vía se basa en una selección de materiales representativos. Y la selección es el enfoque utilizado por los archivistas, bibliotecarios y profesionales del patrimonio para identificar los materiales que formarán las colecciones. En esta se sugiere considerar como criterios el tema, creador/procedencia y formato.[169] Aun cuando las recomendaciones de la Unesco tienen un carácter general formulan un llamado de atención en torno a la selección como tarea inicial en la creación del patrimonio digital.

168 Unesco, *Directrices Unesco/PERSIST* sobre selección del patrimonio digital para su conservación a largo plazo, 2.ª ed., mayo de 2021.

169 *Idem*.

Lo cierto es que la selección puede ser tan amplia y abarcar todo un dominio o tan limitada que solo se refiera a un conjunto específico de recursos de la Web.[170] La selección es la vía necesaria para asegurar la salvaguarda de al menos una parte del patrimonio digital representado por las páginas web y las redes sociales.

Desde finales de la década de los 2010 han surgido varios estudios que profundizan sobre la valoración y selección en el Archivado Web y de redes sociales incorporando las perspectivas teórico-metodológicas de la archivología. Dentro de la perspectiva archivística la valoración se entiende como una función fundamental, donde se determina qué documentos deben ser conservados. En términos teóricos, la valoración ha sido debatida por décadas, particularmente a partir de los 70. Como explica Foscarini, es a partir de estos años cuando archivistas comienzan a cuestionar de manera más abierta los métodos de valoración documental empleados, evaluando su adecuación,[171] una mirada crítica influenciada en buena parte por el posmodernismo, así como por los cambios en las tecnologías de información. Sobre esto último, Foscarini explica que las redes de comunicación digital impactaron significativamente las formas de trabajo y las características de los documentos y los sistemas de información, contribuyendo a su vez a polemizar fundamentos de la teoría y prácticas archivísticas.[172]

Por su parte, Terry Cook, analizando los impactos de los documentos electrónicos en la valoración, ofrece unas reflexiones que pueden ser aplicables a los archivos web y de redes sociales. Cook apunta a la complejidad de las estructuras de los documentos electrónicos, identificándolos como objetos compuestos, y explica lo siguiente:

> El documento "real" por sí mismo no es otra cosa que una imagen fugaz en la pantalla. Además, ese "documento" sólo se puede recrear

170 "About Archiving", International Internet Preservation Consortium.

171 Fiorella Foscarini, "Archival Appraisal in Four Paradigms", en *Currents of Archival Thinking*, 2.ª ed., editado por Heather MacNeil y Terry Eastwood (Libraries Unlimited, ABC-CLIO, 2017), 114.

172 *Idem*.

> con dificultad, pues los datos secundarios o atributivos en los que se basa están continuamente cambiando y el *software* codificado y los enlaces son técnicamente difíciles de recuperar.[173]

El Archivado Web y de redes sociales representa estas características que menciona Cook. Tomando como ejemplo el caso de Twitter, al extraer los tuits de su ambiente natural, estos se almacenan como datos en un fichero JSON, se les aplican métodos computarizados de análisis de datos y de visualización, donde se recrean distintas representaciones de los mismos. De forma similar, los formatos WARC o WARCZ almacenan los contenidos de la web archivada como datos estructurados, donde también se pueden aplicar métodos de análisis de datos y visualización, así como la representación del aspecto visual y funcionalidades de la página web al momento de su acopio. Esto nos invita a considerar las posibles salidas de representación de los contenidos en la Web y en redes sociales como criterio para la selección.

¿Cómo se relacionan estas nociones de valoración archivística al acopio de tuits? Primero, debemos reconocer que en los criterios de selección y en el acopio realizado por las herramientas no se recopila todo. Verne Harris, escribiendo desde el contexto del trabajo de la Comisión de Verdad y Reconciliación de Sudáfrica, enfatiza que los archivos conservan fragmentos de las experiencias sociales.[174] Harris hizo esta reflexión en el contexto de la destrucción sistemática de documentos del régimen apartheid en Sudáfrica, pero podemos pensarlo también en el contexto de archivos de la Web, con el detalle de que se insertan relaciones humanas-tecnológicas que inciden en lo que finalmente se preserva de la Web. El estudio de Summers y Punzalan sobre prácticas de selección en proyectos de archivos web profundiza más sobre este tema, explicando que el Archivado Web abarca condiciones tecnológicas que no son

173 Terry Cook, "Mente sobre la materia: hacia una nueva teoría de la valoración archivística", *Revista d'Arxius*, n.º 3 (2004): 122.

174 Verne Harris, "The Archival Sliver: Power, Memory and Archives in South Africa", *Archival Science* 2, n.º 1-2 (2002).

consideradas en la teoría archivística sobre valoración documental.[175] En otro estudio, Summers enfatiza sobre estos elementos tecnológicos, explicando que los procesos de selección requieren de la incorporación de métodos computarizados que a través del acopio buscan preservar las múltiples dimensiones de los contenidos web.[176]

En resumen, es importante reconocer que, similar a procesos archivísticos, la conformación de archivos web y de redes sociales requiere del establecimiento de criterios de selección, reconociendo que no todo puede ser preservado. Una herramienta importante para establecer estos criterios es a través de una política de curaduría y selección.

4.4. Política de curaduría y selección

El desarrollo de una política de curaduría y selección es el primer paso para poner en marcha un programa de archivado web y de redes sociales. En esta tarea se ha de tomar en consideración el contexto de la institución u organización, el método más adecuado para la formación de colecciones y el criterio de selección.[177]

El contexto de la institución u organización que pone en marcha esta tarea significa considerar la misión, visión, marco legal y actividades de preservación, el personal capacitado y a disposición para llevar a cabo esta tarea, los derechos de autor de las páginas que se pretende acopiar, el equipo técnico (*hardware y software*), los recursos económicos a disposición, así como las posibilidades de uso y aprovechamiento de las colecciones en el futuro.

El método de selección se refiere al modo en que se hará el acopio y recolección. Brown identificó tres métodos: no selectivo, temático y selectivo.[178] El método no selectivo se refiere a la decisión de preservar

175 Ed Summers y Ricardo Punzalan, "Bots, Seeds and People: Web Archives as Infrastructure", en *CSCW'17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (New York, USA: Association for Computing Machinery, 2017), 825

176 Ed Summers, "Appraisal Talk in Web Archives", *Archivaria* 89, n.º 1 (2020): 74.

177 Brown, *Archiving Websites*.

178 *Idem*.

todo sin seleccionar. Esta fue una de las primeras iniciativas del Archivado Web. De hecho, Internet Archive basó su filosofía de archivado en esta idea. En la actualidad este método es poco práctico porque como se ha mencionado con antelación, la publicación de páginas web parece una tarea infinita. Solo en el caso de las instituciones públicas o empresas que deciden guardar toda la memoria de sus publicaciones web, se pueden establecer flujos de trabajo para que el programador de la página guarde una copia de las páginas como testimonio y memoria institucional. Esto significa involucrar al personal del área de tecnologías de la información en una tarea de preservación digital.

El método temático delimita el alcance del Archivado Web. Se trata de un proceso costoso, implica tiempo, es una acción subjetiva e intervienen perspectivas personales, aun cuando se tomen en consideración los lineamientos institucionales previamente establecidos. Puede realizarse por materia, creador, género, dominio.[179]

El método selectivo se identifica con un alto nivel de detalle de las páginas que deben ser preservadas. Un ejemplo de esta práctica fue puesto en marcha por la Biblioteca Nacional de Australia en el Programa PANDORA. Este tipo de selección es de utilidad en la gestión de derechos de autor. También facilita el entendimiento de las propiedades y características de las colecciones digitales.

Los criterios de selección establecen los cánones que delimitan el alcance del acopio. A continuación, se enumeran algunos criterios que pueden ser aplicados:

1) **Tema:** se puede determinar el acopio de páginas web de acuerdo con el contenido que puede ser de carácter social, científico, política, medioambiental, artístico, entre otros.
2) **Derechos de autor:** la usabilidad y reaprovechamiento de los datos es un rasgo significativo en la selección de los materiales que se van a seleccionar. En muchas ocasiones se carece de los

179 *Idem.*

derechos de uso para poner las colecciones en acceso abierto, sin embargo, no se puede descartar su importancia como un material para la investigación científica y la docencia.

3) **Extensión:** la naturaleza de la Web como un documento hipertextual incide en que su preservación sea una tarea compleja. La incorporación de los vínculos, dificulta el alcance y extensión de la página web. Por ello, resulta relevante determinar qué tipo de contenidos y cuál es el alcance de las páginas que se acopiarán. Este aspecto será revisado con más detalle en el siguiente capítulo referido al acopio y recolección.
4) **Tiempo y periodicidad o frecuencia de la publicación:** en algunos casos las páginas que no se actualizan pueden tener programas o aplicaciones obsoletas. Considerar estas problemáticas es un buen punto de partida en el momento de formular una propuesta de preservación. Para determinar qué materiales acopiar es importante considerar con qué frecuencia se edita o modifica la información. Así como determinar la fecha en que se hará la recopilación de la información. Como se ha mencionado con antelación, hay páginas estáticas que no se actualizan y también hay publicaciones, por ejemplo, de medios de comunicación que de forma constante editan nuevos datos.
5) **Determinación de riesgo de desaparición de la publicación:** en esta variable se consideran páginas que por su naturaleza y por el tipo de información que publican pueden desaparecer de un momento a otro. Un ejemplo de este tipo puede ser el conflicto armado o la guerra.
6) **Tópico y significado:** el valor que la información publicada puede tener para un grupo social, comunidad, país o región es un elemento a considerar.
7) **Rareza:** se refiere a publicaciones inéditas y originales que precisamente por este carácter conviene guardar como testimonio de una época. Este tipo de acopios sobre todo corresponden a

centros de investigación o bibliotecas y archivos destinados a rastrear este tipo de contenidos.

Los programas de archivado web y redes sociales operan como un ciclo continuo de selección y creación de colecciones. Esta es una actividad intelectual que es subjetiva porque se identifican y seleccionan materiales que se preservarán para el futuro y cuya usabilidad y pertinencia documental será la base para la creación de colecciones digitales.

Las instituciones y organizaciones interesadas en establecer programas de archivado web y de redes sociales deberán diseñar políticas de curaduría y selección como punto de partida para crear sus colecciones. Esta tarea compete inclusive a las instituciones de alcance nacional como son las bibliotecas y archivos nacionales que podrían estar interesadas en delimitar los rastreos de sus páginas web a través criterios de curaduría y selección. En cada caso, en función del contexto, se deberá definir el método y establecer los criterios de selección.

4.5. ¿Quiénes pueden preservar la Web y las redes sociales?

En las instituciones de la memoria que preservan colecciones analógicas que han digitalizado y que, además, acopian materiales de origen digital, el tránsito al Archivado Web y de redes sociales debiera ser un proceso de evolución natural. Sin embargo, a veces esta ruta no es evidente. Las motivaciones para emprender programas de archivado web y de redes sociales son diversas. De acuerdo con los resultados de un estudio puesto en marcha por Costa, Gomes y Silva[180] la mayor parte de las iniciativas de archivado de este tipo de materiales se dedican a preservar páginas de su país, región e institución. Y solo tres iniciativas se ocupan de archivar materiales de todo el mundo: Internet Archive, Portuguese Web Archive e Internet Memory Fundation.

180 Costa, Gomes y Silva, "La evolución del archivo web", 191-205.

El Archivado Web y de redes sociales es una práctica encaminada a la salvaguarda de la memoria digital que puede ser puesta en marcha por diferentes instituciones, organizaciones e incluso personas, como por ejemplo investigadores interesados en crear colecciones específicas. A continuación, se enuncian algunas de las razones para poner en marcha esta forma de archivado.

Para las instituciones públicas el Archivado Web puede ser el medio para preservar sus publicaciones *online*, en correspondencia con el mandato de acceso a la información, transparencia y rendición de cuentas; para salvaguardar campañas públicas, comunicados de prensa y proyectos especiales, con el fin de contar con un testimonio de las acciones de gobierno, como una evidencia legal, y así generar material para la investigación científica y periodística, entre otros.

Para una empresa puede significar la oportunidad de documentar la historia de la compañía; conservar los materiales promocionales y campañas publicitarias y cumplir con obligaciones de transparencia y acceso a la información, entre otras.

En tanto, para una organización no gubernamental es una tarea a través de la cual se narra la historia de la organización, se ofrece como un recurso para la investigación científica y periodística y se documentan las aportaciones sociales.

Los intereses de una biblioteca o archivo son diferentes. Las bibliotecas y archivos nacionales se ocupan de cosechar los dominios de cada país para cumplir con el depósito legal. Además, en colaboración con otras bibliotecas y archivos pueden promover la creación de colecciones sobre eventos o sucesos específicos, como por ejemplo, los procesos electorales, pandemias, guerras, desastres naturales, entre otros. O bien, proponer la creación de colecciones de comunidades por región, zona cultural, lengua, identidad étnica, demandas de grupos sociales, entre otras. Incluso se puede promover la creación de colecciones de música, sitios turísticos, creaciones artísticas, por citar solo algunas.

Para los museos, la salvaguarda de este tipo de materiales representa la oportunidad de documentar las exposiciones, los programas de difusión y de conservar los materiales sonoros, audiovisuales y textuales complementarios a las exposiciones.

Las fonotecas de radio y las videotecas de televisión pueden incorporar a sus tareas convencionales de preservación la salvaguarda de las páginas web, porque estas constituyen una modalidad de medio digital donde, además de las transmisiones convencionales, se incorporan otro tipo de contenidos complementarios, como por ejemplo, blogs específicos, foros o sitios creados a propósito de ciertos eventos periodísticos de relevancia noticiosa.

En el caso de las universidades, los criterios de selección pueden provenir de los temas de investigación o de actualidad que a los investigadores y docentes les interese recuperar. Por su parte, las bibliotecas de centros e institutos de investigación científica tienen ante sí la posibilidad de crear colecciones digitales de acuerdo con intereses de investigación científica. Son susceptibles de ser preservados *websites*, cuentas de redes sociales, sets de datos, citaciones de artículos publicadas en la Web, páginas Github, foros y plataformas de colaboración, entre otras. Otro uso potencial del Archivado Web es la preservación de obras de arte en Internet y contenido web interactivo como, por ejemplo, obras de arte que han sido digitalizadas, catálogos, cómics en línea, juegos en línea, etcétera.[181]

181 "About Archiving", International Internet Preservation Consortium.

V. Acopio y cosecha de datos

5.1. La cosecha de datos

El acopio de documentos es la primera tarea para la formación de fondos y colecciones en archivos y bibliotecas. Los materiales se ingresan, compran, reciben en depósito, donan o entregan mediante comodato y son ingresados para su resguardo y tratamiento documental. Una vez que los materiales se incorporan a la institución para su salvaguarda, son sometidos a procesos de inventario, limpieza, estabilización, registro en la base de datos, asignación de metadatos y conservación.

El acopio de las páginas web y de set de datos de redes sociales como documentos nativos digitales es distinto. El profesional de la información, sea bibliotecario, archivista, documentalista o curador, crea y cuida el desarrollo de las colecciones digitales. A diferencia de los procesos documentales aplicados a colecciones formadas por soportes físicos como son, por ejemplo, libros, videocasetes, cintas de carrete abierto, discos, entre otros, con los documentos nativos digitales se llevan a cabo actividades de rastreo, es decir, de seguimiento e identificación de la información web, así como de acopio, recolección o captura de información para crear fondos y colecciones digitales.

El acopio en el contexto digital tiene además implicaciones en los roles de los profesionales de la información, dado que se requiere una intervención más activa en los procesos de conformación de colecciones. En otras palabras, no es suficiente esperar a que los materiales digitales lleguen a la institución para ser tratados. Por lo tanto, el profesional de la información asume un rol de agente social activo en la conformación de colecciones digitales, particularmente cuando se trabaja con objetos nativos digitales.

El seguimiento y acopio conforman la fase técnica mediante la cual se identifican y recopilan los datos de los sitios web, a través del uso de un *software* (rastreador web) que descarga, en función de una serie de parámetros y alcances previamente establecidos, el código, imágenes, sonidos, textos, metadatos y otros archivos esenciales para poder reproducir de forma completa y fiel el sitio web.[182]

En el caso de las redes sociales, este proceso de rastreo y recolección se puede realizar de dos maneras: a través del proceso de *web crawling* que se utiliza para el acopio de publicaciones de la Web o a través del acopio de los datos estructurados por medio de la interfaz de programación de aplicaciones (API por sus siglas en inglés).

La información puede ser rastreada y recolectada de forma manual o automatizada. En inglés, se utilizan *crawling*, rastrear y *harvesting* que significa cosechar para nombrar a este proceso. La idea que proponen estos términos se refiere a indagar o seguir el rastro de una o varias publicaciones de la Web y de redes sociales para ser cosechadas en función de los criterios y frecuencia de captura establecidos en la Política Curatorial y de Selección. Esta es una tarea intelectual y práctica.

5.2. Procesos de acopio

El acopio de información publicada en la Web se realiza mediante rastreadores (*crawler*), también llamados arañas y *bots* o *robots* virtuales

182 *Idem*.

que son algoritmos utilizados para analizar el código de un sitio web. Los *crawler* fueron empleados en un principio solo con propósitos de indización. No obstante, su uso se extendió en Internet de tal forma que se utilizan para realizar actividades de seguimiento de los sistemas informáticos para encontrar debilidades, comparar sitios web, eliminar enlaces inactivos, investigar el volumen de búsqueda de palabras clave, eliminar errores ortográficos y archivar páginas. Desde hace tres décadas los rastreadores son herramientas que se emplean en el Archivado Web.[183]

Se rastrean o buscan las páginas, se copia la información en un archivo WARC y se habilita su conservación y acceso digital. Para ello, se consideran los criterios de selección y la lista de direcciones URL (Uniform Resource Locator).

> Dependiendo de la arquitectura del servidor web y del nivel de interacción con el cliente, los rastreadores pueden capturar el sitio web completo o sólo algunas partes del mismo. La parte que queda fuera del alcance de los rastreadores se ha denominado «Web profunda» o «Web oculta» en la terminología de los motores de búsqueda.[184]

El rastreo de la Web puede variar en función de la dimensión y nivel de profundidad de acopio, pero hay algunas características comunes en este proceso:

1) Programación e inicio del rastreo

Los rastreadores desarrollan su actividad a partir de una indicación inicial. Se establecen de forma automatizada las direcciones URL que deben buscar o bien, se define de forma manual la navegación personalizada de la Web que se desea capturar.

Para realizar rastreos masivos se definen puntos originales de acopio

183 *Idem*.

184 Masanès, "Web Archiving", 22

que se denominan semillas.[185] Lo que significa que se determina el listado de páginas web que formarán las colecciones digitales. De esta forma el rastreador:

> (...) las analiza, extrae los enlaces y recupera el documento enlazado. Repite este proceso con el documento obtenido y continúa mientras tenga enlaces que explorar y encuentre un documento dentro del ámbito definido. Este proceso es necesario, ya que el HTTP no proporciona un comando que devuelva la lista completa de documentos disponibles en el servidor (...). Por lo tanto, cada página tiene que ser "descubierta" mediante la extracción de enlaces de otras páginas.[186]

Las recolecciones masivas de la Web las realizan robots de rastreo (*web crawlers*) que recorren Internet e identifican los contenidos que serán recopilados con base en un listado inicial de URL o semillas. Los robots guardan de forma automática los sitios identificados:

> (...) recorriendo todos los hiperenlaces que encuentran y añadiéndolos a la lista de URL que visitarán recursivamente, almacenando de forma ordenada los documentos, videos, imágenes, etcétera, que se ocultan tras cada vínculo. Estos contenidos son indexados posteriormente con el fin de poder realizar las búsquedas a texto completo en la colección.[187]

Ilustra este proceso la experiencia de la Biblioteca Nacional de España, institución que desde 2009 comenzó a recolectar sitios web españoles identificados con el dominio .es; así como otros dominios y subdominios genéricos (.com; .edu; .gob; .org; .net; etcétera). "Las recolecciones buscan reproducir con detalle el aspecto del sitio y las funcionalidades del mismo disponibles durante la captura, de tal forma que la réplica del sitio web sea tan navegable como su versión *viva*".[188]

185 "About Archiving", International Internet Preservation Consortium.

186 Masanès, "Web Archiving", 23.

187 María Jesús Molina Suárez, "Archivo web de las publicaciones en línea en las comunidades autónomas", *Cartas diferentes. Revista canaria de patrimonio documental*, n.º 15 (2019): 290.

188 "Archivo de la Web española/Historia de la colección", BNE, consultado el 16 de enero de 2023, https://www.bne.es/es/colecciones/archivo-web-espanola.

El rastreo masivo de las páginas web que serán cosechadas por la Biblioteca Nacional de España inicia con la ingesta de la información mediante la cual se programará en el CWEB (Colecciones en la Web) la búsqueda de sitios.

La CWEB es una herramienta desarrollada por la Biblioteca Nacional de Francia y utilizada por la Biblioteca Nacional de España, en el marco del convenio de colaboración de las dos instituciones con el International Internet Preservation Consortium (IIPC).[189] En este desarrollo informático los conservadores o curadores digitales trabajan "introduciendo datos como URL, parámetros de recolección (frecuencia, profundidad y tamaño) y descripción (materia, palabras clave, notas de contenido y notas técnicas)".[190] Mediante este proceso se establecen las listas de los sitios que deberán ser rastreados. Esta información se incorpora después en el NAS (NetarchiveSuite) paquete de herramientas utilizadas para el manejo de los robots. De esta forma se despliega la cosecha masiva de la Web mediante estos procesos automatizados de acopio y recolección masiva.

Por su parte, el rastreo manual de sitios web se utiliza sobre todo en la creación de colecciones digitales con fines de investigación y docencia. En este procedimiento no se sigue una lista de URL. Se captura la información a medida que navega por la Web y se elige la información que deberá ser resguardada.

Para este tipo de acopio se pueden utilizar herramientas como Webrecorder, cuya aplicación Archiveweb.page se incrusta en navegadores como Chrome. Así, desde el navegador se recolecta la información de la página y los vínculos seleccionados. Este rastreador permite el acopio de todo tipo de información (textual, sonora, audiovisual, etcétera). Los materiales acopiados en formato *WARC* o *WARCZ* pueden almacenarse, hasta cinco GB, de forma temporal en Webrecorder. Las limitaciones en la captura están determinadas por el nivel de acceso que ofrece el creador del sitio. Un ejemplo de este tipo de rastreo corresponde a la

189 Molina, "Archivo web de las publicaciones".
190 *Ibid*., 291.

colección que sobre COVID-19 en México puso en marcha el grupo de investigadores del proyecto PAPIIT IT 400121 del Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM.

2) Inserción de Robots.txt y limitaciones

La inserción de *Robots.txt*, protocolos incrustados en un sitio web, indica a los rastreadores que no capturen el sitio web o que recojan solo partes de él. Impiden el archivado de la página o de algunas partes del mismo.[191] Si durante el archivado de la Web, el curador se encuentra con estos casos, podría establecer comunicación con los creadores de las páginas para solicitar permisos para el archivado de la información. Este tipo de acciones se ponen en marcha sobre todo cuando se necesita preservar por la encomienda de Depósito Legal alguna página web.

Otra consideración a tomar en cuenta durante el acopio es la limitación que impone el protocolo HTTP para proporcionar una copia masiva del contenido del servidor. "Los servidores HTTP sólo pueden entregar su contenido archivo por archivo, siempre que se solicite su URL. Esto hace que el descubrimiento de la ubicación de cada archivo sea una de las cuestiones clave en el Archivado Web".[192]

3) Profundidad de rastreo o granularidad de la información

Durante el proceso de acopio se debe determinar el nivel de granularidad o profundidad de recogida de la información.[193] Los rastreadores se configuran para acopiar un determinado número de enlaces. Es decir, el nivel y especificidad de la información que se cosechará. De acuerdo con la tecnología utilizada para tal fin, se puede recopilar una página o bien navegar y guardar la información publicada en todos o en determinados *links*.

Como se ha señalado antes, la información publicada en la Web es hipermedia. Esto significa que se guardan contenidos en todo tipo de

191 "About Archiving", International Internet Preservation Consortium.
192 Masanès, "Web Archiving", 21.
193 Hanna, "El Modelo".

lenguajes: textuales, audibles, audiovisuales y fotografías. Bajo esta consideración, el acopio conlleva la toma continua de decisiones en relación con el tipo de contenido, los *links* o hipervínculos navegables, así como la cantidad y calidad de la información que será recogida y que podrá ser consultada a largo plazo por parte de los usuarios actuales y del futuro.

El curador determina la extensión y calidad del documento web y de redes sociales que recuperará. Se procura recuperar la mayor cantidad de información posible para que el contenido capturado tenga relevancia documental y uso potencial en el futuro.

4) Frecuencia y duración

La frecuencia de cosecha de las páginas web es una decisión que ha de ser establecida en las *Políticas Curatoriales y de Selección* y puesta en marcha durante el proceso de acopio y rastreo. En la definición de la frecuencia de rastreo se ha de considerar la naturaleza del sitio y los ciclos de actualización de la Web.

Lo anterior significa que el curador debe conocer la frecuencia de actualización de los sitios seleccionados para ser acopiados. Para comprender este aspecto, deben anotarse como ejemplo que las páginas web de medios de comunicación se actualizan varias veces al día, en función de los sucesos periodísticos; en tanto que las páginas de muchas instituciones públicas cuentan con información fija que salvo contadas ocasiones se modifica.

La duración empleada para el acopio de una página web puede ser de unos cuantos segundos o minutos; o hasta emplear horas o incluso días, dependiendo del volumen de la información que se requiera acopiar. En cada situación la perspectiva es diferente dado que existen una serie de factores que además de los antes señalados, también deben ser tomados en consideración, entre los que destacan, la verificación de la fuente de información, inserción de aplicaciones para evitar la incorporación de mensajes publicitarios en las páginas acopiadas, restricciones de acceso a la información publicada, contenidos protegidos con contraseña,

materiales que sólo pueden ser consultados en una búsqueda local dentro de un sitio web.

En algunos casos este proceso puede ser complejo y lento. En ocasiones el equipo de cómputo mediante el cual se realiza el rastreo personalizado puede tener problemas de conexión y falta de espacio para el procesamiento y almacenamiento de los archivos. Por ello, se ignoran archivos pesados como son los de audio y video, porque la descarga de este tipo de materiales puede prolongarse durante mucho tiempo. Esta visión práctica contrasta con la encomienda de crear documentos únicos, cuya permanencia es incierta dado que muchos materiales se publican de forma temporal en Internet.

El acopio y la recolección son procesos intelectuales y técnicos complejos y exhaustivos, en los cuales el curador requiere de habilidades en el manejo de la tecnología mediante la cual programará o emprenderá la captura de la Web y de la información publicada en redes sociales. Este proceso requiere del análisis de la calidad de información, la pertinencia documental y de una perspectiva objetiva y neutral durante la captura de contenidos. Asimismo, se deberá tomar en cuenta la naturaleza hipermedia y colectiva de la Web para guiar la captura y extenderla tanto como sea posible. La meta del archivado web sería entonces cuantitativa. Y desde esta perspectiva han trabajado varias bibliotecas nacionales y el Archivo de Internet a escala mundial. La experiencia en el Archivado Web y de redes sociales en las últimas tres décadas testifica que las iniciativas aisladas son insuficientes porque no pueden proporcionar por sí solas la extensión, profundidad y calidad de los contenidos archivados. Al final de cuentas, como señala Masanès "Los distintos esfuerzos se considerarán parte de un único archivo global cuando la interconexión entre los archivos web se organice… Sólo así los usuarios podrán aprovechar todos estos esfuerzos y obtener la mejor memoria web posible".[194]

194 Masanès, "Web Archiving", 20.

5.3. Tipos de acopio y recopilación

La web no es un archivo simple. La estructura de este tipo documental es compleja y en cada publicación diferente. Por ello, la estrategia que deben adoptar archivistas y bibliotecarios en cada situación debe ser diferente. De acuerdo con el contexto de cada institución u organización que desee emprender un programa de archivado web y de redes sociales se derivan cuatro tipos de acopio y recopilación.

Recopilación transversal y cosecha masiva de sitios web

Esta modalidad es empleada sobre todo por bibliotecas y archivos nacionales que tienen a su cargo el Depósito Legal de la Web. Este tipo de acopio se dispone a partir del rastreo y recopilación instantánea de los dominios establecidos. La frecuencia para realizar la cosecha masiva puede variar en función de cada institución. Por ejemplo, realizan el acopio amplio o masivo una vez al año la Biblioteca Nacional de España que cosecha los dominios .es;[195] el UK Web Archive que recopila los sitios web .uk, .scot, .wales, .cymry y .london para cumplir con el depósito legal de Reino Unido[196] y la Biblioteca Nacional de Francia que acopia sitios franceses.[197] En tanto, la Biblioteca Nacional de Dinamarca lleva a cabo capturas masivas de publicaciones web danesas,[198] cuatro veces al año.

Recopilación selectiva y sistemática de sitios web

Además del acopio masivo, se realizan cosechas con un mayor nivel de profundidad y frecuencia. Esta modalidad es una opción para las instituciones de la memoria que tienen la obligación de salvaguardar el

195 "Archivo de la Web española", BNE, https://bnelab.bne.es/dato/archivo-de-la-web-espanola/.

196 "Frequently asked questions", UKWA UK Web Archives, consultado el 19 de enero de 2023, https://www.webarchive.org.uk/en/ukwa/info/faq/#how-frequently-are-websites-collected.

197 "Consulter les Archives de l'internet", BnF, consultado el 19 de enero de 2023, https://www.bnf.fr/fr/archives-de-linternet.

198 "Netakivet", Det Kgl. Bibliotek, consultado el 19 de enero de 2023, https://www.kb.dk/en/find-materials/collections/netarkivet.

Depósito Legal y que por su exhaustividad resulta insuficiente la captura anual. Por ello, se pueden poner en marcha recolecciones temáticas varias veces al año. Se incluyen, en este modo, por ejemplo, el acopio de medios de comunicación, partidos políticos, organizaciones y asociaciones, ministerios o secretarías de Estado y agencias, perfiles seleccionados de las redes sociales, videos de YouTube, Tik Tok, entre otros. La recopilación selectiva y sistemática se emplea también para crear colecciones digitales con fines de investigación y docencia. Las colecciones especiales se pueden desarrollar como resultado de una petición de investigación o bien como una decisión curatorial por considerar que en un futuro la información capturada tendrá valor patrimonial, como recurso de información, testimonio o como un insumo para la creación de nuevos contenidos. En la tabla 1 se ofrecen algunos ejemplos de colecciones temáticas en biliotecas nacionales.

Tabla 1

Biblioteca	Temas
Biblioteca Nacional de España[199]	Medioambiente y cambio climático, Bellas artes y cartografía, Feminismo, Organismos públicos.
UK Web Archive[200]	British Stand-Up Comedy, French in London, History, Politics and Government; Art and Culture; Places, Society and Communities.
Biblioteca Nacional de Corea[201]	Cambio climático, 70 aniversario de la Guerra de Corea, Contenido cultural, Violencia deportiva.
Library of Congress[202]	Official Campaign Website - Tonya Lynn Millis; Connecticut Democratic Party; State of North Dakota; Partido Independentista Puertorriqueño (PIP); The American Nazi Party.

Fuente: Elaboración propia con información de las bibliotecas y archivos.

199 "Archivo de la Web española", BNE.

200 "Topics and Themes", UKWA UK Web Archives, https://www.webarchive.org.uk/en/ukwa/category/.

201 "OASIS (Online Archiving & Searching Internet Sources)", National Library of Korea OASIS, consultado el 19 de enero de 2023, https://nl.go.kr/oasis/.

202 "Format Web Archive", Library of Congress, https://www.loc.gov/web-archives/.

Recopilación de acontecimientos de emergencia e interés social

El Archivado Web y de redes sociales es un proceso dinámico y creativo. Lejos de ser repetitivo y monótono se funda en el acopio de información de relevancia social que amerita ser preservada. Dado que la Web es un espejo del pensamiento y la creación de la sociedad del siglo XXI, es el medio a través del cual se narra y edifica la historia contemporánea y se desarrollan algunos de los hechos y debates más grotescos, insólitos y valiosos de la humanidad. Son ejemplo de ello, los procesos electorales, los desastres naturales, las epidemias, las guerras, entre otros. La atención e intención del Archivado Web debe ser la actualidad; en consecuencia, el procesamiento documental de la Web es una tarea de permanente actualización de las colecciones.

Gracias a esta perspectiva se han creado sendas colecciones sin las cuales no podríamos comprender a la sociedad del siglo XXI. Por ejemplo, la colección *Japan Earthquake* que salvaguardó información en relación con el terremoto que asoló al país nipón en 2011 y que derivado de las circunstancias del desastre el Archivado de la Web se desarrolló como iniciativa encabezada por la Universidad de Virginia.[203] Otra colección de relevancia para la humanidad es *Saving Ukrainian Cultural Heritage Online* (*SUCHO*) forjada como resultado de una iniciativa mundial, en la que participaron más de mil voluntarios en tareas de archivado web, para salvaguardar la herencia digital cultural de Ukrania ante la invasión Rusa.[204]

Acopio desde el servidor, mediante el Protocolo de Transferencia de Archivos (FTP) o transferencia de archivos en un dispositivo electrónico

Para el acopio desde el servidor se requiere el permiso del programador o editor de la misma. El acopio mediante FTP o la transferencia de

203 Lori Donovan, "Japan Disaster Archives: Collaboration for successful web archiving", Archive It, publicado el 28 de febrero de 2013, https://archive-it.org/blog/post/japan-disaster-archives-collaboration-for-successful-web-archiving/.

204 "About SUCHO", SUCHO, consultado el 7 de marzo de 2023, https://www.sucho.org/about.

archivos son vías empleadas por bibliotecas nacionales que deben resguardar las publicaciones por mandato de Depósito Legal.

5.4 Acopio y recopilación de contenidos de redes sociales en el ambiente API

Como se explica en el capítulo IV, el acopio de contenidos de redes sociales se puede llevar a cabo de dos maneras. Una estrategia es a través del proceso de *web crawling* o de captura, tal como se realiza para el acopio de sitios y páginas web. Esta estrategia se conoce como la captura del *look and feel*. La segunda estrategia es la de acopio a través de la API de las plataformas de redes sociales. Esta modalidad se conoce como la de acopio de datos estructurados y se ha utilizado principalmente para el acopio de tuits. La decisión sobre la estrategia a seguir dependerá en gran medida de lo que se desea preservar.

Si el interés del proyecto es recrear el estilo visual y la funcionalidad de la página al momento de la captura, se utiliza la estrategia de captura. Si el interés principal es el acopio de los datos de los contenidos en redes sociales para aplicar métodos computacionales de análisis y visualización se utiliza la estrategia de acopio de datos estructurados.

Es importante mencionar otras características del acopio de datos estructurados que se distinguen de la estrategia de captura *look and feel*. Justin Littman menciona las siguientes:[205]

- El acopio por medio de la API permite mayor flexibilidad para establecer parámetros de búsqueda y recuperación. Por ejemplo, se puede realizar acopio por etiquetas o por usuario.
- El acopio por medio de la API permite recopilar metadatos que no son posibles a través de la captura de páginas web.
- Un desafío del acopio por medio de la API es que cada plataforma

205 Justin Littman, "Web archiving and/or/vs social media API archiving", Social Feed Manager, publicado el 13, de diciembre de 2017, https://gwu-libraries.github.io/sfm-ui/posts/2017-12-13-web-social-media-archiving.

de redes sociales utiliza su respectivo API y por lo tanto se requiere de diferentes *softwares* para el acopio.

- Relacionado con el punto anterior, no todas las plataformas de redes sociales proveen acceso abierto a la API. Además, los términos de uso varían.

Conviene recordar que la API consiste en una serie de protocolos utilizados para comunicar *softwares* entre sí con el propósito de consultar datos, analizar respuestas y enviar instrucciones. La figura 1 presenta una representación general del funcionamiento de la API. En esta representación el cliente, a través de un código de programación, envía instrucciones a la API para recopilar o analizar datos almacenados en otra plataforma.

Figura 1

Fuente: Elaboración propia.

Aplicando este proceso al acopio de tuits, el mismo se fundamenta en lograr acceso a la API de Twitter y utilizar un programa al que se le da una instrucción para que la API identifique y recupere tuits basados en los parámetros de búsqueda establecidos en la instrucción. La API acopia los tuits y los almacena en un archivo formato JSON. La figura 2 muestra la representación de este proceso.

Figura 2

Fuente: Elaboración propia.

En este ejemplo, se desea realizar el acopio de tuits que contengan la etiqueta #covid19mx. La API de Twitter recibe la instrucción y accede al archivo de Twitter, identificando y recopilando los tuits con dicha etiqueta y generando como respuesta a la instrucción un archivo en formato JSON con todos los tuits recopilados.

Aunque en términos generales esta representación del proceso de acopio por medio de la API puede ser aplicable a diversas plataformas de redes sociales, las particularidades de cada plataforma, incluyendo sus políticas de acceso y uso de los datos, impactan el archivado de sus contenidos para propósitos de preservación. Hasta el año 2023, el acceso a la API de Twitter incluía una opción gratuita para propósitos de investigación académica. Es por esta razón que la estrategia de acopio por medio de la API ha sido utilizada principalmente para Twitter. Bibliotecas y archivos han utilizado la estrategia de captura del *look and feel* para otras plataformas como Facebook e Instagram.

En términos generales, el acopio de contenidos de Twitter por medio de la API tiene las siguientes características:

1) **Identificar los parámetros para la búsqueda y recopilación de tuits**
   Este paso está vinculado al proceso de selección que se explica en el capítulo IV. En el contexto de Twitter, se evalúa si se va a realizar un acopio temático a partir de la identificación de etiquetas utilizadas por los usuarios, a partir de términos y/o enfocados en usuarios.

2) **Obtener acceso a la API de Twitter**
   Para lograr acceso a la API de Twitter es necesario registrarse en el portal para desarrolladores de aplicaciones en Twitter (https://developer.twitter.com) y crear un proyecto donde se solicita el acceso a la API. Una vez aprobada la solicitud, el proyecto genera una serie de claves llamadas: clave del consumidor, secreto del

consumidor, *access token* y secreto del *access token*. Estas claves se utilizan en la configuración del programa seleccionado para el acopio, las cuales permiten obtener el acceso a la API. Es importante mencionar que Twitter establece límites de acopio de tuits por mes. Además, en marzo de 2023, la empresa anunció un sistema de pago para lograr acceso a la API. Antes de este anuncio el acceso era gratuito para la investigación académica y sólo requería de la aprobación de la solicitud. Este nuevo sistema divide el acceso en tres niveles, de los cuales uno es gratuito, pero su límite de tuits por mes que pueden acopiarse es de solo 1 500.[206]

3) **Utilizar una herramienta de acopio de tuits por medio del API de Twitter**

Se han desarrollado múltiples herramientas para realizar acopio de tuits por medio del API de Twitter (ver la sección 5.5). Una de las más utilizadas por bibliotecarios y archivistas es *twarc*, una librería de líneas de comando en lenguaje Python desarrollado por la organización Documenting the Now.[207] El siguiente ejemplo, basado en el acopio de tuits con la etiqueta #covid19mx, ilustra el proceso de acopio por medio de la API de Twitter utilizando *twarc*.

La línea de comandos básica para realizar el acopio es 'twarc search':
twarc search #covid19mx > tweets_covidmx_20221111.jsonl
donde 'twarc search´ es la instrucción que se da a la API de Twitter, #covid19mx es el parámetro de búsqueda y tweets_covidmx_20221111.jsonl es el fichero que se creará con el set de datos de tuits recopilados.

206 "Twitter's v2 API", Developer Portal, consultado el 29 de marzo de 2023, https://developer.twitter.com/en/portal/products/free.

207 "Documenting the Now", DN, consultado el 29 de marzo de 2023, https://www.docnow.io.

5.5. Herramientas tecnológicas para el acopio de la web y de las redes sociales

Desde hace tres décadas ha evolucionado el *software* destinado a rastrear y acopiar páginas web. La primera herramienta que se desarrolló fue Wget en 1996. Esta aplicación se hizo popular dos años después, cuando se distribuyó en GNU/Linux. Un año después, *Internet Archive* ofreció como herramienta *open source* Heritrix, uno de los rastreadores más utilizados en bibliotecas y archivos de alcance nacional que recopilan grandes volúmenes de contenidos. Desde entonces y hasta el primer semestre de 2023 se han desarrollado más de 15 *softwares open source* y servicios para llevar a cabo cosechas masivas y acopio personalizado de la Web y redes sociales. Se infiere que el desarrollo de las herramientas ha alcanzado un estado de maduración y que su uso se incrementa de forma significativa.

De todos los desarrollos informáticos destinados al rastreo y copiado de la Web y de redes sociales destacan Heritrix, el primer rastreador que puso en marcha Internet Archive. A partir de este se han impulsado otros desarrollos. El uso de Heritrix destaca en los rastreos masivos y es empleado por bibliotecas nacionales. Aunque conviene señalar que también puede ser utilizado para el acopio de colecciones más pequeñas.

Este *software* se adapta para la cosecha de colecciones formadas por miles de semillas URL. Esta aplicación puede configurarse para que no se capture material idéntico. Acopia los contenidos en formato WARC, aunque en un principio utilizó ARC. Produce archivos web que reflejan las mejores prácticas y estándares de archivado web. En 2014, se desarrolló Umbra, un complemento que mejora el desempeño de Heritrix en la captura de contenidos complejos y dinámicos en la Web.

En cuanto al acopio y recolección manual de la Web destaca Webrecorder, cuyo antecedente fue Conifer, creada en 2014 por Ilya Kreymer, con el apoyo de la Fundación Andrew W. Mellon. En 2020, su creador anunció el cambio de nominación de Conifer a Webrecorder.[208]

208 Ilya Kreymer, "A New Phase for Webrecorder Project, Conifer and ReplayWeb.page", Webrecorder Web archiving for all!, publicado el 11 de junio de 2020, https://webrecorder.net/2020/06/11/webrecorder-conifer-and-replayweb-page.html.

Webrecorder consta de varias herramientas a través de las cuales se lleva a cabo el rastreo y acopio de la web de forma manual. Webrecorder puede ser útil en la creación de colecciones digitales en centros de investigación, bibliotecas y archivos interesados en desarrollar esta modalidad de archivado. En 2021, Webrecorder propuso el formato WARCZ para que el archivado fuera una tarea más sencilla de realizar.

Además de las tecnologías, se han puesto en marcha servicios, de estos destaca Archive-It, empleado en el inicio de programas de archivado web y redes sociales por diversas bibliotecas, archivos e instituciones de la memoria. Este servicio fue creado desde 2006 para proveer de herramientas, capacitación y soporte técnico para la captura, almacenamiento y preservación de páginas web; así como para ofrecer una plataforma que se gestionara de forma independiente a Internet Archive para compartir y dar acceso a las colecciones. Hasta el primer semestre de 2023, Archive-It ofrece servicio a más de 800 organizaciones de 24 países.[209] En la tabla 2 se puede apreciar la evolución de los *softwares* desarrollados para el rastreo y acopio de la Web.

Tabla 2. Software y servicios para el Archivado Web y de redes sociales

Año de creación	Herramienta	Características
1996	Wget[210]	Herramienta gratuita y de código abierto bajo la Licencia Pública General GNU. Está disponible sólo en línea de comandos para descargar archivos de la Web. Está diseñada para ejecutarse en entornos Unix y similares (incluyendo Mac OS), pero también tiene una versión para Windows. Soportará capturas HTTP, HTTPS y FTP.

209 "About Archive-It", Archive-It, consultado el 30 de marzo de 2023, https://archive-it.org/learn-more/.

210 "GNU's Wget", GNU Operating System, consultado el 9 de junio de 2023, https://www.gnu.org/software/wget/.

Año de creación	Herramienta	Características
1997	Heritrix[211]	Cuenta con licencia *open-source* y fue programado en JAVA. Rastreador (o *crawler*) de ficheros web a través de Internet. Disponible para todas las versiones de Windows y entorno Unix.
1998	HTTrack[212]	*Software* libre y de código abierto que utiliza la licencia GPL para descargar en HTML un sitio WWW desde Internet a un directorio local. WinHTTrack es la versión de Windows (de Windows 2000 a Windows 10 y superior) de HTTrack, y WebHTTrack es la versión de Linux/Unix/BSD.
2000	NWA Toolset	Creada con PHP, Perl y Java. Utiliza estándares abiertos como el protocolo HTTP y XML para la comunicación entre diferentes partes del sistema. El código fuente de NWA Toolset está bajo la Licencia Pública General GNU (GPL). Desde enero de 2004, NWA Toolset ha sido un proyecto en SourceForge.
2003	WERA (Web Archive Access)	WERA es una aplicación PHP basada en Nutchwax. Reemplaza a NwaToolset. WERA incluye una aplicación web *arcretriever* para obtener registros de directorios de archivos ARC de Internet Archive.
2004	Netarchive Suite[213]	Su desarrollo se basó en Heritirix. Es *software* libre y está disponible bajo la licencia LGPL.
2006	Archive-It[214]	Es una herramienta desarrollada por Internet Archive para ayudar a las organizaciones en el acopio, cosecha y creación de colecciones web.

211 "Home", Internet Archive/Heritrix3, consultado el 9 de junio de 2023, https://github.com/internetarchive/heritrix3/wiki.

212 "Bienvenido", HTTrack WEBSITE COPIER, consultado el 9 de junio de 2023, https://www.httrack.com/.

213 "About", NetarchiveSuite/NetarchiveSuite, consultado el 9 de junio de 2023, https://github.com/netarchivesuite/netarchivesuite#readme.

214 Archive-It, consultado el 9 de junio de 2023, https://archive-it.org/.

Año de creación	Herramienta	Características
2007	Web curator tool[215]	Herramienta para la recopilación selectiva de la Web. Está integrada con la versión 3 de Heritrix. Fue diseñada para ser utilizada por bibliotecarios y no necesariamente personal técnico.
2017	WAIL (Web Archiving Integration Layer)[216]	Es una aplicación de escritorio programada en Python que integra Heritrix y OpenWayback.
2013	Perma.cc[217]	Servicio de Archivado Web creado por el laboratorio de innovación de la Universidad de Harvard.
2014 2020	Conifer- Webrecorder	ArchiveWeb.page se incorpora como una extensión en Chrome para el acopio de la Web.
2014	Heritrix+Umbra[218]	Soporta la captura de JavaScript y permite un desplazamiento dinámico. Umbra es una herramienta de automatización para archivado web que se ejecuta junto con Heritrix. Imita la forma en que un navegador accede a la página.
2015	Grab-a-site[219]	Rastreador web de fácil configuración diseñado para realizar copias de seguridad de sitios web.
2015	Twarc[220]	Librería de líneas de comando en el lenguaje de programación Python que extrae tuits por medio de la API de Twitter.

215 Web Curator Tool, "Overview and History", wct, https://webcuratortool.readthedocs.io/en/latest/guides/overview-history.html#introduction.

216 John A. Berlin *et al.*, "WAIL: Collection-Based Personal Web Archiving", en 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, editado por IEEE (2017), 1-2. DOI: 10.1109/JCDL.2017.7991619.

217 "About Perma.cc", Perma.cc, consultado el 9 de junio de 2023, https://perma.cc/

218 "Umbra", Internet Archive/Umbra, consultado el 9 de junio de 2023, https://github.com/internetarchive/umbra.

219 "Grab-site", ArchiveTeam/grab-site, consultado el 10 de junio de 2023, https://github.com/ArchiveTeam/grab-site.

220 "Twarc", DN Twarc, consultado el 10 de junio de 2023, https://twarc-project.readthedocs.io/en/latest/.

Año de creación	Herramienta	Características
2016	Social Feed Manager[221]	*Software* de código abierto que recopila datos de redes sociales. Esta herramienta fue creada por las Bibliotecas de la Universidad George Washington. Captura contenidos y metadatos de Twitter, Flickr y Sina Weibo.

Fuente: Elaboración propia con información de las páginas de los desarrollos.

En el contexto de Twitter, cabe destacar Twarc, una librería de líneas de comando en el lenguaje de programación Python que extrae tuits por medio de la API.[222] La herramienta fue desarrollada en el 2014 por Ed Summers, desarrollador de *software* especialista en curaduría, preservación digital y sustentabilidad.[223] Summers desarrolló twarc para realizar acopio de tuits sobre las protestas en Ferguson, Missouri, contra el abuso policiaco hacia comunidades negras y otros grupos minoritarios y por la muerte del joven afroamericano Michael Brown por parte del policía blanco Darren Wilson.[224] Cinco días después del asesinato de Brown, Summers participaba como ponente en una mesa de la conferencia anual del Society of American Archivists que activó una conversación sobre el rol de los archivistas en documentar este suceso.[225] Al siguiente día Summers se enteró de la iniciativa de Internet Archive para crear un archivo web, detonando su reconocimiento sobre la importancia de también realizar el archivado de tuits.[226]

221 Social Feed Manager, consultado el 9 de junio de 2023, https://gwu-libraries.github.io/sfm-ui/.

222 "Twarc", DN twarc.

223 "Social.coop", Mastodon, https://social.coop/@edsu.

224 Para más información sobre la muerte de Michael Brown y las protestas en Ferguson ver: "El asesinato de Michael Brown y los disturbios raciales en Ferguson: todas las claves", 20minutos, 19 de agosto de 2014, https://www.20minutos.es/noticia/2217996/0/claves-asesinato-michael-brown/disturbios-raciales-eeuu/ferguson-misuri/; "Amnistía Internacional publica un nuevo informe sobre Ferguson que documenta abusos contra los derechos humanos", Amnistía Internacional España, publicado el 24 de octubre de 2014, https://www.es.amnesty.org/en-que-estamos/noticias/noticia/articulo/amnistia-internacional-publica-un-nuevo-informe-sobre-ferguson-que-documenta-abusos-contra-los-derec/.

225 Ed Summers, "A Ferguson Twitter Archive", Inkdroid, publicado el 30 de agosto de 2014, https://inkdroid.org/2014/08/30/a-ferguson-twitter-archive/.

226 *Idem*.

A partir del proyecto de acopio de tuits sobre las protestas en Ferguson, *twarc* se convirtió en una de las herramientas de acopio de tuits más utilizadas por bibliotecarios, archivistas, e investigadores. Forma parte de una serie de herramientas desarrolladas por la iniciativa Documenting the Now para el acopio y curaduría de tuits.[227] El Catálogo DocNow provee acceso a los identificadores de tuits de 144 colecciones temáticas, cubriendo tuits del periodo 2006 a 2021.[228] Aunque *twarc* funciona a base de líneas de comando en lenguaje Python no se requiere de conocimiento extensivo en programación. La comunidad de prácticas que utiliza *twarc* ha elaborado extensa documentación sobre las funcionalidades y uso de *twarc*.[229]

La decisión de eliminar el acceso gratuito a la API de Twitter para propósitos de investigación académica ha impactado tanto a investigadores como a quienes desarrollan herramientas que requieren interactuar con la API. El 18 de abril de 2023 Documenting the Now informó que Twitter revocó las claves de la herramienta Hydrator, utilizada para recuperar tuits públicos de los identificadores de tuits acopiados, y la cual estuvo funcionando por siete años.[230] También fueron revocadas las claves de DocNow, una herramienta diseñada para incorporar los pasos vinculados al proceso de acopio de tuits y publicación de identificadores.[231] Al momento de escribir este libro, todavía existen preguntas sobre cómo incorporar prácticas de archivado de redes sociales que no dependan del acceso a la API. Ante esta realidad, la comunidad de Documenting the Now propone que se considere la implementación de estrategias de captura utilizando herramientas de archivado web como

227 "Documenting the Now", DN.

228 "Catalog", DN, consultado el 5 de marzo de 2023, https://catalog.docnow.io.

229 "Twarc", DN Twarc.

230 DocumentingTheNow [@documentnow], "The Hydrator app has recently had its keys revoked. The extremely limited read access in Twitter's new API quotas mean it will no longer work. Hydrator has operated for the last 7 years to help users reconstitute tweet ID datasets, but no more...", Tweet, publicado el 18 de abril de 2023, https://twitter.com/documentnow/status/1648325488236961792?s=20.

231 Ed Summers, "Looking Forward", Medium, publicado el 19 de julio de 2023, https://news.docnow.io/looking-forwards-64cee8436640.

Webrecorder, así como colaborar directamente con creadores de contenido para que donen sus archivos de cuentas de redes sociales.[232] Esta realidad puede abrir mayores oportunidades para desarrollar proyectos colaborativos que involucren distintos actores sociales, incluyendo los propios creadores de contenido.

232 *Idem.*

VI. Gestión y almacenamiento digital

6.1. Desarrollos *open source* para la gestión y búsqueda de información

La preservación digital de la Web y de la información publicada en redes sociales comprende todas las actividades y estrategias necesarias para garantizar la integridad, autenticidad y accesibilidad de los datos. En este libro se han anotado los procesos a partir de los cuales se puede emprender un programa de archivado web y de redes sociales. Se establecieron la curaduría y la selección como el primer binomio de tareas intelectuales para iniciar de este modo el trabajo documental. Después, se añadieron el acopio y el rastreo como procesos técnicos para capturar información. En este capítulo se abordarán dos procesos indispensables para garantizar la permanencia de la información digital: la gestión y el almacenamiento.

Los gestores de contenidos son las herramientas a través de las cuales se administran y manejan sendos volúmenes de datos. Su advenimiento y notabilidad se asocia con la necesidad de administrar los objetos digitales y metadatos obtenidos como resultado de la digitalización de soportes analógicos.

Antes de que se usaran los gestores de contenidos digitales se emplearon las primeras bases de datos que, desde la década de los 70 del siglo

pasado, se programaron para organizar y manejar, a través de la administración de los metadatos, las colecciones analógicas en bibliotecas y archivos. Después, en la década de los 90, cuando se emprendieron los procesos de digitalización de materiales analógicos, se pusieron en marcha los primeros sistemas de gestión y almacenamiento masivo digital para manejar, además de los metadatos, los contenidos digitales.

Estos sistemas se insertaron como parte del ciclo de vida digital y fueron empleados para relacionar los procesos de ingesta, el establecimiento del ID para vincular al objeto analógico con el digital, la identificación e incorporación de metadatos derivados de la catalogación, el almacenamiento, la búsqueda y la recuperación de contenidos.

El gestor de contenidos es un componente del archivo digital sustentable. Con base en el Open Archival Information System (OAIS) puede ser definido como el conjunto de servicios y funciones necesarias para generar, mantener y hacer accesibles los contenidos, la información descriptiva –que identifica y documenta los fondos y colecciones–, los datos administrativos del catálogo y el registro estadístico de los datos contenidos; así como para dar acceso a los contenidos.[233]

En el caso del Archivado Web y de redes sociales, los gestores digitales son las aplicaciones o *softwares* a través de los cuales se organizan fondos y colecciones, se identifican los materiales recopilados, se conservan y se proporciona acceso a los datos. Los datos de las páginas web y de redes sociales se consultan mediante una plataforma por medio de la cual se realizan búsquedas mediante la URL, palabras o colecciones. También se facilita la presentación y difusión de las colecciones. Además, cuentan con el software con el cual se pueden visualizar, escuchar y leer los datos acopiados en los formatos WARC y JSON. Algunas de las herramientas para la gestión y visualización de la Web archivada son: Wayback Machine, Open WayBack, Python Wayback (PYWB), SolrWayback,

233 Perla Olivia Rodríguez Reséndiz, Joséphine Simonnot y Dafne Citalli Abad Martínez, "Gestor de contenidos de código abierto para archivos digitales sonoros que preservan materiales de investigación", *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32, n.º 77 (2018): 101-115. http://dx.doi.org/10.22201/iibi.24488321xe.2018.77.58005.

Webrecorder Player, Browsertrix Cloud[234] y Archipelago. Además, se han creado desarrollos recientes a través de los cuales se procesan grandes volúmenes de datos archivados mediante técnicas computacionales para generar nuevas colecciones y conocimiento. La inserción de la Inteligencia Artificial (IA) es una modalidad de trabajo en crecimiento. A continuación, se enuncian algunas experiencias notables.

Wayback Machine

WayBack Machine es una aplicación creada en 2001 por Internet Archive para capturar, almacenar, gestionar y dar acceso a páginas web. Para los científicos es una valiosa fuente de información. En 2003, después de dos años de acceso público, Wayback Machine creció a un ritmo de 12 terabytes por mes. Los datos se almacenan en sistemas de PetaBox diseñados por el personal de Internet Archive. El primer rack de 100 TB entró en pleno funcionamiento en 2004.[235]

En 2005, se desarrolló Open WayBack. Este desarrollo en Java fue creado para contar con una aplicación pública que mejore las funcionalidades de WayBack Machine. OpenWayback admite dos modos de acceso o reproducción: URL de archivo y Proxy.[236] La Biblioteca Nacional de España (BNE) utilizó hasta 2019 Open WayBack, para la recuperación de la Web mediante la URL. En la actualidad, la BNE desarrolla otras herramientas para hacer búsquedas por palabras, sin embargo, la cantidad de datos es tal que esta no es una tarea sencilla. Otra alternativa propuesta es la recuperación por colecciones o títulos.[237]

Con el apoyo del International Internet Preservation Consortium (IIPC), Ilya Kramer desarrolló Pywb (Python Wayback) un kit de he-

234 International Internet Preservation Consortium, "Tools and Software", IIPC, https://netpreserve.org/web-archiving/tools-and-software/.

235 "Internet Archive Wayback Machine", Internet Archive, consultado el 9 de mayo de 2023, https://archive.org/web/.

236 Tyler A. Young, "General overview", IIPC/OpenWayback, editado el 24 de julio de 2018, https://github.com/iipc/openwayback/wiki/General-overview.

237 Molina, "Archivo web de las publicaciones", 283-307.

rramientas para que los usuarios que lo requieran puedan instalar una WayBack Machine personal.[238]

Archive-It

En el 2006 el Internet Archive lanzó Archive-It, una plataforma de archivado web que incorpora las diferentes fases de la gestión de archivos web, desde la selección y captura al acceso. En términos de captura, Archive-It aplica la estrategia de web *crawling*, donde se le provee a la plataforma una lista de URL de los contenidos que se desean acopiar para conformar la colección, además se especifica la frecuencia y profundidad de captura. La plataforma permite además crear metadatos a nivel de colección y de ítem utilizando el esquema Dublin Core, y la publicación de la colección. El proyecto de Archivado Web "Novel coronavirus outbreak", del IIPC, utilizó Archive-It como el gestor digital para la conformación, publicación y preservación de los contenidos web seleccionados.[239]

El servicio de Archive-It requiere del pago de una licencia anual que depende del espacio de almacenamiento que se requiera, y del tipo de plan. Estos planes son en dólares estadounidenses, lo cual es una limitación. Sin embargo, el Internet Archive incluye otros programas que consisten en colaboraciones con organizaciones con recursos limitados, como por ejemplo Archive-It Sponsored.[240] Otro programa es Spontaneous Event Collections, donde el Internet Archive colabora con organizaciones e individuos para crear una colección web sobre un evento en específico.[241] En este programa, la organización que propone la creación de una colección web sobre un evento realiza el trabajo de selección de

238 "A wayback machine (pywb) on a cheap, shared host", Literary Machines, publicado el 24 de octubre de 2014, https://literarymachin.es/pywb-wayback-machine/.

239 International Internet Preservation Consortium, "Novel Coronavirus (COVID-19)", Archive-It, archivado desde febrero de 2020, https://archive-it.org/home/IIPC y https://archive-it.org/collections/13529.

240 "Archive-It Sponsored", Archive-It, consultado el 7 de marzo de 2023, https://archive-it.org/blog/archive-it-sponsored/.

241 "Spontaneous Event Collections", Archive-It, consultado el 8 de marzo de 2023, https://archive-it.org/blog/spontaneous-events/.

contenidos y creación de metadatos, mientras que el grupo de trabajo de Archive-It administra la plataforma realizando el acopio de los contenidos e importando los metadatos. El Proyecto #RickyRenuncia, una iniciativa para documentar las protestas que llevaron a la renuncia del gobernador de Puerto Rico en julio de 2019, colaboró con Archive-It por medio de este programa para crear la Colección Web #RickyRenuncia.[242]

SolrWayback

SolrWayback es una aplicación para la gestión y búsqueda de archivos web. Esta herramienta fue desarrollada por la Biblioteca Real Danesa. El SolrWayback se instala en un servidor Solr con archivos ARC/WARC indexados con el indexador WARC. Entre otras funcionalidades permite búsquedas de texto, enlaces interactivos en los gráficos de los sitios, exportación de los resultados de búsqueda en WARC, búsquedas de imágenes similares. La Biblioteca Nacional Széchényi de Hungría puso en línea una demostración para SolrWayback https://webadmin.oszk.hu/solrwayback/.[243]

Browsertrix Cloud

Browsertrix Cloud es un desarrollo open *source* de WebRecorder creado en 2022, para programar y dar seguimiento de forma automatizada al rastreo, gestión y almacenamiento de páginas web en la nube.[244] Fue diseñado para ejecutarse en el entorno de la nube en plataformas

242 Internet Archive Global Events, "#RickyRenuncia web collection (Puerto Rico 2019)", Archive-It, archivado desde julio de 2019, https://archive-it.org/collections/12491. Para conocer sobre el Proyecto #RickyRenuncia y el proceso de conformación de la colección web ver: Joel Antonio Blanco-Rivera, Irmarie Fraticelli Rodríguez y Marisol Ramos, "Documentando lo espontáneo: las protestas #RickyRenuncia", *Archidata: Boletín de la Red de Archivos de Puerto Rico* 18, n.º 1 (2020): 13-17.

243 National Széchényi Library of Hungary, "SolrWayback", consultado el 10 de marzo de 2023, https://webadmin.oszk.hu/solrwayback/.

244 "Automated Browser-Based Crawling at Scale", Browsertrix Cloud, consultado el 10 de marzo de 2023, https://browsertrix.cloud/features/.

como Kubernetes, aunque cada usuario puede elegir el soporte de almacenamiento que más les convenga de acuerdo con su infraestructura informática.

Archipelago Commons

En 2022, auspiciado por el Metropolitan New York Library Council, se lanzó Archipelago Commons, que de forma simplificada es Archipelago, un desarrollo informático *open source* para crear archivos o repositorios que preservan una amplia diversidad de objetos digitales.[245] Dentro de los cuales destacan las páginas web y los datos de redes sociales.

Proporciona en *open source* la arquitectura de un servidor DAM (Digital Asset Management), para crear un repositorio de objetos digitales basado en CMS Drupal 8/9 y publicado bajo GPL V.3 en evolución basada en el popular CMS Drupal 8/9 y publicada bajo la licencia GPL V.3 License.[246]

Archipelago fue creado para ser utilizado por la comunidad de Galleries, Libraries, Archives and Museums (GLAM). Es decir, por los profesionales que preservan contenidos en diversas instituciones de la memoria. A través de este gestor se describen, almacenan, vinculan, exponen metadatos y contenidos en una amplia gama de formatos.[247]

Admite la incorporación de los siguientes tipos de objetos digitales: artículos, libros, set de datos, documentos digitales, mapas, películas, pinturas, fotografías, tesis, páginas, *podcasts*, objetos 3D, páginas web, objetos de audio, objetos de video, pósteres, periódicos manuscritos, entre otros.

Cada uno de estos objetos se incorpora en el formato digital en el cual fue creado, por ejemplo, WARC para páginas web; JSON para set de

245 "Archipelago Commons Intro", Archipelago Documentation, última modificación el 18 de octubre de 2021, https://docs.archipelago.nyc/1.0.0/.

246 "Archipelago Commons Documentation Repository", Esmero/Archipelago-Documentation, consultado el 10 de marzo de 2023, https://github.com/esmero/archipelago-documentation.

247 *Idem*.

datos de redes sociales; WAVE para audio o bien PNG, GIF, JPG o JPEG para imágenes o bien PDF para textos, entre otros. Admite el acopio de objetos individuales o bien colecciones de objetos digitales. Es decir, que este gestor puede preservar tanto materiales que son resultado del proceso de digitalización como aquellos cuyo origen es digital, como son las páginas web y las redes sociales.

6.2. Almacenamiento versus conservación

Se emplea el término almacenar para referirse a la acción de guardar o archivar algo. Esta palabra se ocupa de forma recurrente en el ámbito archivístico, aunque en muchas ocasiones su aplicación es equívoca. Por ejemplo, se señala que los materiales de archivo se guardan en el almacén. Quienes desconocen el trabajo documental se refieren con esta palabra a que los materiales permanecen guardados y en muchas ocasiones hacinados durante varios años o incluso décadas.

Se nombra con la palabra almacén a las bóvedas, es decir, a los espacios destinados dentro del archivo para conservar los materiales documentales. Las bóvedas son los lugares designados a la conservación de los documentos en adecuadas condiciones de temperatura, humedad y limpieza, de los documentos. La conservación es el proceso documental a través del cual se posibilita que los contenidos registrados en cualquier tipo de soporte permanezcan y puedan ser consultados a largo plazo.

Conservar es proteger por medio de técnicas y tecnologías un documento original para que no pierda su integridad[248] y prolongar su vida útil en el soporte original en el cual fue creado. A través de las técnicas de conservación preventiva se monitorean y previenen los posibles deterioros de los materiales. En tanto que con la conservación correctiva se subsanan o al menos se intentan revertir los posibles deterioros del documento.

248 Ray Edmondson, *Filosofía y principios de los archivos audiovisuales*, 1.ª ed. en español (Unesco, IIBI-UNAM, UASLP, 2018), 103.

A finales del siglo pasado, con la irrupción de la tecnología digital, se advirtió que la conservación era una tarea insuficiente para preservar y dar acceso, sobre todo a los materiales grabados en soportes analógicos. La transferencia de contenidos a soportes digitales fue avalada como la única vía que existe hasta ahora para que los contenidos puedan ser consultados y reaprovechados[249] en el ecosistema digital actual.

La digitalización es, desde entonces, una tarea titánica a través de la cual se produjeron objetos digitales con la información registrada en diversos soportes físicos. La acumulación de este tipo de contenidos devino en nuevas complejidades como el manejo y resguardo de un creciente volumen de contenidos digitales.

Cuando iniciaba el siglo, en la Conferencia Anual de la Asociación Internacional de Archivos Sonoros y Audiovisuales (IASA), el ingeniero alemán Albrecht Haefner señaló el surgimiento de una nueva tendencia teórica y tecnológica que ganaba reconocimiento en los archivos de la radio y la televisión: el almacenamiento masivo digital.[250]

Con este término se denominó al espacio físico y virtual donde se guardaban los objetos digitales derivados de la digitalización y también aquellos cuyo origen es digital. También se usa la noción bóvedas digitales para referirse al *hardware*, donde se salvaguardan, organizan y conservan los objetos digitales. Con ello, se incorporó la terminología del archivo clásico a su versión digital.

En el almacenamiento masivo digital se preservan grandes volúmenes de datos a través de métodos adecuados de conservación. La conservación es una parte de la preservación digital a través de la cual se ponen en marcha métodos y tecnologías para que los ítems digitales no pierdan las propiedades de autenticidad, integridad y usabilidad.

---

249 Perla Olivia Rodríguez Reséndiz, "La preservación digital sonora". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 30, n.º 68 (2016): 173-195, https://doi.org/10.1016/j.ibbai.2016.02.009.

250 *Idem*.

6.3. Principios para mantener la integridad de los datos

Derivado del acopio y rastreo de páginas web y de redes sociales se crean paquetes de datos en formatos de preservación como WARC y WARCZ, y formato de datos como JSON. Estos son documentos nativos digitales, también son conocidos como objetos digitales que poseen tres propiedades:[251] sus datos se conservan en un medio o soporte físico digital; su información es interpretada y procesada de forma lógica mediante un *software* y su contenido es reconocido y comprendido por las personas mediante una aplicación de computadora.[252] Las propiedades de los documentos como objetos digitales y su estrecha vinculación y dependencia con el *software* y *hardware* son los factores de la obsolescencia o entrada en desuso de la tecnología, uno de los mayores problemas que confronta la preservación digital.

"El almacenamiento digital es el medio para la conservación, administración y gestión de objetos digitales".[253] Los soportes digitales más utilizados para el almacenamiento digital son discos duros, servidores, la nube y discos Linear Tape Open (LTO). La capacidad del volumen de almacenamiento y la durabilidad de los soportes es diferente en cada caso.

Los principales deterioros que pueden tener los soportes digitales son:

Obsolescencia:

Los dispositivos de almacenamiento pueden tornarse obsoletos cuando dejan de ser compatibles con el *hardware* al que se conectan. Frente a la entrada en desuso de los soportes digitales se han sugerido estrategias

251 Michael Day, "The Long.Term Preservation of Web Content", en *Web Archiving*, editado por Julien Masanès (Berlin: Springer, 2006), 177-194.

252 Kennet Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years", en *The State of Digital Preservation: An International Perspective* (Washington, DC: Council on Library and Information Resources, 2002), 4–31. http://www.clir.org/pubs/abstract/pub107abst.html y https://www.clir.org/wp-content/uploads/sites/6/pub107.pdf.

253 Rodríguez Reséndiz, *Estado de la preservación digital*, 64.

de preservación digital encaminadas a garantizar la permanencia de los objetos digitales. De estas destacan dos: la migración y la emulación. La más utilizada es la migración sistemática de un soporte que progresivamente entra en desuso por otro reciente. Se estima que el tiempo de vida útil de los soportes digitales no excede de los cinco años. Por lo tanto, es necesario contar con un plan de migración cíclico.

Otra estrategia de preservación digital es la emulación de paquetes de *software* para que los objetos digitales se puedan ver en el *software* original en el que fueron creados, incluso en una computadora actual. Esta estrategia es compleja porque requiere de contar con personal especializado para emular programas obsoletos.

Pérdida de los bits:

El tiempo de vida de un soporte digital puede variar entre los cuatro o cinco años. Al paso de este tiempo se pueden presentar fallas que dañen los archivos.

Los errores humanos:

Sea de forma intencionada o no, el inadecuado manejo de los dispositivos puede ser una amenaza latente para los archivos.

Los desastres naturales o catástrofes sociales:

Los fenómenos naturales como son las inundaciones, huracanes e incendios son amenazas potenciales que ponen en riesgo los datos. Y, además, las guerras, saqueos o robos también pueden dañar o hacer que se pierdan los contenidos digitales.

Falta de recursos económicos:

La carencia de un presupuesto asignado de forma sistemática afecta en el mantenimiento de los sistemas de almacenamiento masivo digital.

Los sistemas de almacenamiento masivo digital deben considerar los siguientes principios para garantizar la integridad y autenticidad de los datos:[254]

Redundancia y diversidad. Utilizar soportes de almacenamiento digital heterogéneos de tal forma que, si uno de ellos presenta fallas, la información pueda ser recuperada de otro. Realizar al menos dos copias del material digital y ubicarlas en lugares que estén geográficamente alejados. Se sugiere tener almacenamiento *online, offline* y *near offline*. Se ofrece la consulta de los materiales *online*, en general se proporciona acceso a copias en baja resolución. *Near offline* es la modalidad para tener acceso a contenidos de uso frecuente y cuyo acceso necesita ser inmediato. Esta modalidad se emplea en medios que necesitan contar con la información de manera inmediata. También se recomienda contar con almacenamiento *offline* en soportes como son, por ejemplo, las cintas LTO para evitar una indebida manipulación y afectación del soporte.

Verificación seguimiento y reparación. Los procesos de *check sum* o verificación de sumas se emplean cíclicamente para monitorear la integridad de la información. En caso de que se observe corrupción de información se pueden hacer copias de reemplazo. En esta tarea es de utilidad mantener las copias de almacenamiento en lugares separados.

Vigilancia de la evolución tecnológica y evaluación de riesgos. Como una medida para contar con acciones proactivas que mitiguen problemas que afecten los materiales y además, posibiliten la creación de prospectiva de tecnologías que se utilizarán a futuro.

Documentación del almacenamiento y simplicidad. En la medida en que se puedan documentar las acciones de conservación y almacenamiento masivo digital será más sencilla la preservación digital, será una

254 Digital Preservation Coalition, "Storage", DPC, https://www.dpconline.org/handbook/organisational-activities/storage.

tarea que se mantenga a lo largo del tiempo. Además, es conveniente que se simplifiquen los medios y sistemas de almacenamiento.

6.4. Experiencias de gestión y almacenamiento masivo digital

Internet Archive

El crecimiento exponencial en cuanto a almacenamiento digital de páginas web es imparable.

> Los archivos de la World Wide Web conservaron de 1996 a 2010 un total de 181 978 millones de contenidos (6,6 PB). Internet Archive por sí solo contenía 150 000 millones de contenidos (5,5 PB). En 2014, todas las iniciativas habían archivado juntos al menos 534 604 millones de contenidos, lo que suma alrededor de 17 PB de datos. Esto representa un aumento de 2010 a 2014 del 294 % en el contenido y del 258 % en el volumen de datos.[255]

Internet Archive es el más grande archivo de páginas web en el mundo. Su creador Brewster Kahle, inspirado en la legendaria Biblioteca de Alejandría y en la novela *Moby Dick* de Herman Melville, fundó un proyecto grande que no terminará y que durará más allá de su vida.[256] Las instalaciones de Internet Archive, donde se cuenta con un espacio de almacenamiento de 100 petabytes, están situadas en una iglesia histórica en Richmond en San Francisco California.[257]

Internet Archive comenzó en 1996, archivando la propia Web. Conforme avanzó esta iniciativa sus colecciones digitales se diversificaron. En la actualidad ofrece, como se señaló en el capítulo I, el acceso a 41 millones de libros y textos; 14.7 millones de grabaciones de audio (con 240 000

255 Costa, Gomes y Silva, "La evolución del archivo web", 191–205.

256 Ferose VR, "Digital Librarian for and of the World", Medium, publicado el 14 de junio de 2023, https://ferosevr.medium.com/digital-librarian-for-of-the-world-9ec7cf1a239e.

257 *Idem*.

conciertos); 8.4 millones de videos (de los cuales 2.4 son noticias de televisión); 4.4 millones de imágenes y 890 000 programas de *software.*[258]

En relación con las páginas web, en 2002 se cuantificaron 10 000 millones de sitios. En 2016, Internet Archive reportó que almacenaba 273 000 millones de páginas web de 361 millones de *websites*, con un almacenamiento estimado en 15 petabytes.[259] Y hasta el 16 de junio de 2023, almacena 815 000 millones de páginas web.[260] En dos décadas incrementó de manera exponencial el volumen de información almacenado.[261] Internet Archive sigue siendo, con diferencia, el archivo web con mayor volumen de información.

El sistema de almacenamiento digital de Internet Archive se denomina Petabox. En 2021, estaba conformado por cuatro centros de datos, 745 nodos y 28 000 discos giratorios. El almacenamiento digital de las páginas web se resguarda en Wayback Machine con una capacidad de 42 petabytes.[262] El sistema de almacenamiento digital no cuenta con aire acondicionado, en su lugar se utiliza el exceso de calor para ayudar a calentar el edificio.

Internet Archive es posiblemente la mayor biblioteca y archivo con la que haya contado la humanidad. Tiene centros de respaldo en San Francisco, Egipto (en la mítica ciudad de Alejandría), Canadá y en Países Bajos.[263]

WayBack Machine es la herramienta para la gestión y búsqueda de la información acopiada en Internet Archive. A través de esta herramienta se llevan a cabo consultas históricas. Se pueden buscar las páginas a través de la incorporación de la URL, por palabra o bien por colecciones. Además, en la interfaz inicial se pueden visualizar y consultar algunas

258 Internet Archive, "About the Internet Archive".

259 Vinay Goel, "Defining Web pages, Web sites and Web captures",*Internet Archive Blogs*, 23 de octubre de 2016, "https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.

260 "Internet Archive Wayback Machine", Internet Archive.

261 "Internet Archive", Wikipedia, última modificación el 27 de octubre de 2023, https://en.wikipedia.org/wiki/Internet_Archive#cite_note-86.

262 "Petabox", Internet Archive, https://archive.org/web/petabox.php.

263 Ferose VR, "Digital Librarian for".

de las páginas capturadas desde 1996.[264] La consulta de la información se habilita mediante un calendario que presenta en qué fecha fueron capturadas las páginas solicitadas. El usuario elige la página que desea consultar.

En cuanto a Archive-It, el Internet Archive se encarga del almacenamiento digital en sus centros de respaldo de las colecciones manejadas por las instituciones que utilizan este servicio.[265] Las instituciones pueden descargar una copia de sus colecciones para almacenarlas en sus respectivos espacios y gestionarlas en sus propios programas de preservación digital. Sin embargo, la mayoría de las instituciones no descargan su copia, aduciendo principalmente limitaciones de presupuesto y de personal dedicado al archivado web y a la preservación digital.[266]

ARCH (Archives Research Compute Hub)

Internet Archive desarrolló ARCH para los académicos que están desarrollando investigación científica con grandes colecciones de datos en diversos ámbitos científicos, entre otros se pueden señalar, como ejemplo, a las humanidades digitales. En esta plataforma los usuarios crean colecciones personalizadas y acceden a grandes cantidades de datos a través de la minería de datos, ciencia de datos, *machine learning* (ML) entre otros. La ARCH promueve la publicación en acceso abierto y la publicación de grandes conjuntos de datos generados por los usuarios. "ARCH también procesa partes de Wayback Machine que asciende a más de 50 PB y se remonta a 1996, lo que supone un extenso archivo de la historia y la comunicación contemporáneas".[267]

264 "Internet Archive Wayback Machine", Internet Archive, http://web.archive.org/.

265 "About Archive-It", Archive-It.

266 Bailey, entrevistado por Blanco-Rivera.

267 Tpadilla, "Build, Access, Analyze: Introducing ARCH (Archives Research Compute Hub)", *Internet Archive Blogs*, 26 de junio de 2023, https://blog.archive.org/2023/06/26/build-access-analyze-introducing-arch-archives-research-compute-hub/.

IA Copilot

La incorporación de herramientas de Inteligencia Artificial para la búsqueda y recuperación de información en grandes volúmenes de datos es una práctica que se ha popularizado en los últimos años. Algunas herramientas como ChatGPT, Google Bard y HuggingChat utilizan grandes volúmenes de datos como los que ha acopiado Internet Archive para ofrecer respuestas a las preguntas formuladas por las personas. Esta información en muchos casos es errónea. Por ello, una de las polémicas recientes expresa el uso inadecuado por parte de los sistemas de Inteligencia Artificial de la información digital que almacena Internet Archive.[268]

Estas herramientas se entrenan en un periodo de tiempo límite determinado. Son capaces de interpretar las consultas de los usuarios y ofrecer información valiosa. Sin embargo, tienen limitaciones, cuando intentan responder a preguntas que requieren información más allá de la fecha límite de entrenamiento. En estas situaciones, las aplicaciones pueden inventar respuestas que no están respaldadas por los datos disponibles. Para evitar este uso indebido de grandes volúmenes de datos las herramientas deben configurarse para limitar sus respuestas a las fuentes de datos designadas mediante complementos como *Web Request.*

Sumado a lo cual Internet Archive puso en marcha IA Copilot, para interactuar con el contenido guardado en Wayback Machine, utilizando para ello el Chat GPT.[269]

---

268 Ilan K. Cherre, "The Internet Archive está colapsado y la culpa la tiene una IA en proceso de entrenamiento", Computer Hoy, publicado el 29 de mayo de 2023, https://computerhoy.com/internet/internet-archive-colapsado-culpa-tiene-ia-proceso-entrenamiento-1252426.

269 Peter Chan, "Navigating Through Archived Websites: From Text Matching to Generative AI-Enhanced Q&A", International Internet Preservation Consortium, publicado el 28 de junio de 2023, https://netpreserveblog.wordpress.com/2023/06/28/navigating-through-archived-websites-from-text-matching-to-generative-ai-enhanced-qa/.

UK Web Archive

Se documenta que hasta 2023, conservaba 500 TB de información y cada año el crecimiento es de 60 o 70 TB.[270] La información se resguarda en el Digital Library System, repositorio digital creado por la British Library y las Bibliotecas que participan en el Depósito Legal de ese país para la preservación digital a largo plazo.[271] Este sistema mantiene principios para asegurar la permanencia de la información digital. Tiene una red de protección o cortafuegos y protocolos de seguridad contra virus y carece de acceso público a través de Internet. Además, tiene copias espejo de almacenamiento digital de toda la información en cuatro nodos ubicados en St. Pancras, Boston Spa, Aberystwyth y Edinburgh. En estos nodos se verifica de forma constante la integridad de la información, se realizan replicaciones y reparaciones cuando se requiere. Si la información en un nodo comienza a ser corrompida o bien se pierde, de forma automática se restauran los datos de otro de los nodos. Además, cada nodo copia y almacena la información completa en dos o más discos físicos, con autocomprobación y replicación entre los discos.[272]

Biblioteca Nacional de Francia

Por su parte, la Biblioteca Nacional de Francia reporta que almacena más de un petabyte de datos. Las primeras colecciones iniciaron de forma experimental con Internet Archive en 1996. Entre 2007 y 2017, el número de dominios acopiado se incrementó de 0.9 millones a 4.5 millones.[273]

---

270 "Frequently asked questions", UKWA UK Web Archives, https://www.webarchive.org.uk/en/ukwa/info/faq/#how-big-is-the-archive.

271 "Frequently asked questions", UKWA UK Web Archives, https://www.webarchive.org.uk/en/ukwa/info/faq/#where-is-the-archive-stored.

272 "Frequently asked questions", UKWA UK Web Archives, https://www.webarchive.org.uk/en/ukwa/info/faq/#how-big-is-the-archive.

273 "Consulter les Archives de l'internet", BnF.

Biblioteca Nacional de Australia

En 1996, la Biblioteca Nacional de Australia, en colaboración con nueve instituciones de ese país, creó PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) como una iniciativa para la cosecha de la Web. A partir de 2001, puso en marcha el Sistema de Archivo de la Web de Australia, mejor conocido como PANDAS (Preserving and Accessing Networked Documentary Australian System) para gestionar los flujos de trabajo, desde la identificación, selección y cosecha de páginas web hasta la gestión de las restricciones de acceso.[274] La última actualización de este gestor de contenidos digitales se llevó a cabo en 2007.

Web Archiving Project, National Diet Library of Japan

El Proyecto de Archivado Web de la Biblioteca Nacional Diet de Japón comenzó en 2002, como una iniciativa experimental de acopio de algunos sitios web públicos y privados. En 2009, la modificación de la Ley de Bibliotecas posibilitó la cosecha de todos los sitios publicados por instituciones públicas. Desde 2010, realizan una cosecha masiva mensual. El crecimiento de los datos cosechados se ha incrementado de forma significativa, en 2021, alcanzó 2 400 TB de almacenamiento digital. La cosecha de datos se realiza mediante Heritrix, aunque en un inicio utilizaron Wget. Para la gestión, búsqueda y visualización o reproducción de información se utiliza Open Wayback.[275]

274 "PANDORA Digital Archiving System (PANDAS)", Pandora: Australia's Web Archive National Library of Australia and Partners, http://pandora.nla.gov.au/pandas.html.

275 Shimura Tsutomu, "20 Years of the Web Archiving Project (WARP) at the National Diet Library, Japan", International Internet Preservation Consortium, publicado el 3 de abril de 2023, https://netpreserveblog.wordpress.com/2023/04/03/20-years-of-the-web-archiving-project-warp-at-the-national-diet-library-japan/.

Gestión y almacenamiento de archivos de Twitter

Instituciones que han realizado proyectos de acopio de tuits por medio de la API, generando archivos en el formato de datos JSON, han integrado la gestión y almacenamiento de los mismos a sus repositorios digitales. Por ejemplo, la Universidad de Michigan gestiona sets de datos de Twitter sobre temas relacionados con la universidad y el estado de Michigan en su repositorio institucional "Deep Blue Documents" que utiliza DSpace.[276] Por su parte, la Biblioteca Pública de Washington, D. C., utilizó Social Feed Manager para el acopio de tuits sobre el COVID-19 en el distrito,[277] y gestiona los sets de datos en su repositorio digital por medio de ContentDM.[278]

Una razón por la cual instituciones gestionan y almacenan sets de datos de Twitter en los repositorios digitales es el formato de datos JSON. Cada uno de estos archivos en formato JSON puede incluir datos estructurados de cientos de miles de tuits. Por lo tanto, son gestionados como cualquier otro formato de objetos digitales. Además, el enfoque de uso de estos sets de datos es para análisis y visualización, no reproducción. En este sentido, el usuario necesita descargar estos sets de datos para realizar el análisis. Esto es diferente al formato WARC, donde se utiliza un reproductor para que el usuario pueda ver el sitio web capturado.

276 "Twitter Archives", University of Michigan Library, https://deepblue.lib.umich.edu/handle/2027.42/116594.

277 "COVID-19 in Washington, D. C. Twitter Archive", DIGDC, consultado el 9 de mayo de 2023, https://digdc.dclibrary.org/islandora/object/dcplislandora%3A237558?solr_nav%5Bid%5D=7bf56a-3c0a50a9608922&solr_nav%5Bpage%5D=0&solr_nav%5Boffset%5D=1.

278 "Dig DC API Documentation", DC Public Library, consultado el 2 de mayo de 2023, https://dcpubliclibrary.github.io/digdc/.

VII. Acceso y consideraciones éticas

7.1. Búsqueda y consulta de archivos web y de redes sociales

El acceso y reaprovechamiento de los materiales que se salvaguardan en bibliotecas y archivos son actividades de un mismo proceso documental que le confieren sentido a la preservación. Sin la posibilidad de ofrecer para su consulta los materiales que son conservados toda tarea de salvaguarda documental carece de sentido. Como ha señalado Ray Edmondson "la conservación y el acceso son dos lados de la misma moneda".[279]

El acceso es toda forma de uso que se le dé a un documento, sea de naturaleza textual, sonora, audiovisual, fotográfica, multimedia o hipermedia. Los documentos se resguardan para que puedan ser usados y de beneficio para la sociedad.

En la actualidad se busca información sobre todo en Internet. A través de un buscador popular como Google o bien, mediante bibliotecas, archivos y repositorios digitales. La preeminencia del libro como vehículo de conocimiento comparte presencia con documentos sonoros, audiovisuales, fotográficos y con las colecciones de páginas web y con los sets de datos de las redes sociales.

279 Edmondson, *Filosofía y principios*, 27.

Los programas de archivado web y de redes sociales emprendidos, sobre todo en universidades y centros de investigación, centran sus acciones en atender a sus usuarios y comunidades específicas como son los investigadores. El uso potencial de este tipo de materiales en la investigación científica, en las ciencias sociales, historia, sociología, ciencia política, humanidades digitales, entre otras disciplinas, es una tendencia en crecimiento.[280]

En consonancia con lo cual, Emily Reynolds advirtió que desde hace una década se ha recomendado que los archivos web y de redes sociales deben ser interoperables con métodos de investigación como son la visualización, el acceso a través de poderosas herramientas de búsqueda, el análisis de datos sociales y recursos de geolocalización, entre otros.[281] Las posibilidades de estudio e investigación que ofrece la Web y las redes son muy amplias e involucran una amplia gama de disciplinas.[282]

Los científicos constituyen solo un fragmento de un amplio espectro de usuarios que podrían estar interesados en la consulta y uso de este tipo de materiales, cuya naturaleza desafía los modos tradicionales de tratamiento y acceso a la información. No obstante esta relevancia, subsiste un letargo en la incorporación de las prácticas de archivado web y de redes sociales en algunas instituciones de la memoria. En ciertas bibliotecas subsisten las ideas de soslayar, tanto los métodos de búsqueda y navegación -en favor de la catalogación tradicional- como las ayudas para la localización e integración de las colecciones de archivos web, en vez de aprovechar las ventajas de los métodos de descripción y los nuevos sistemas de acceso. La insistencia en adaptar este tipo de materiales de origen digital a categorías bibliográficas, normas o flujos de trabajo que no se crearon teniendo en cuenta los archivos de la Web o las redes

280 Jefferson Bailey *et al.*, *Web Archiving in the United States: A 2016 Survey*, An NDSA Report (National Digital Stewardship Alliance, febrero 2017).

281 Emily Reynolds, *Web Archiving Uses Cases*, Library of Congress, UMSI, ASB13, March 7, 2013, https://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf.

282 Nanna Bonde Thylstrup, "La memoria digital del mundo está en peligro", *The New York Times*, 25 de junio de 2023, https://www.nytimes.com/es/2023/06/25/espanol/opinion/internet-archivo-digital.html?smid=nytcore-ios-share&referringSource=articleShare.

sociales, pueden sofocar una notable oportunidad para la creatividad y la innovación en torno al acceso y la localización.[283]

Las nuevas nociones de búsqueda y acceso a la información a la que están orillando las colecciones de la Web y de redes sociales podrían significar una oportunidad para las instituciones de la memoria, cuyo propósito central es conferir valor de uso a sus fondos y colecciones a través de ensanchar su acceso y reaprovechamiento.

De acuerdo con un estudio publicado en 2016 por la NDSA, prevalecen en instituciones de Estados Unidos que llevan a cabo actividades de Archivado Web y de redes sociales, dos tendencias en el acceso a colecciones de páginas web:

1) Las organizaciones que apoyan la búsqueda a nivel de ítem disminuyó y aumentó la consulta por colecciones como registros de catálogo.[284] Es decir, se evidenció que se confiere notabilidad a las colecciones digitales más que a un ítem específico.
2) En cuanto a las funciones de búsqueda por URL, texto completo, lista de URL y lista por título se mostró una baja en la búsqueda por URL. Esta disminución se vio compensada por el interés mostrado en las colecciones publicadas en las plataformas de acceso.

La visibilidad de las colecciones digitales en las páginas de bibliotecas y archivos podría indicar que éstas adquieren cada vez mayor relevancia y se integran, aunque lentamente, a las tareas tradicionales de preservación, pues para realizar tareas de archivado web y de redes sociales se destina a una sola persona, en el mejor de los casos. Por fortuna, esta actividad va ganando relevancia poco a poco.

En el International Internet Preservation Consortium (IIPC) opera un grupo de trabajo abocado a buscar soluciones comunes a los problemas de acceso. Además de la investigación y desarrollo tecnológico se analizan los aspectos legales, éticos y económicos de este proceso.

283 Bailey *et al*., *Web Archiving in the United States: A 2016 Survey*.
284 *Idem*.

7.2. Consideraciones legales y éticas en torno al acceso

El título de este apartado puede ser entendido como un contrasentido a la idea de acceso abierto a la información que es un principio de toda institución que custodia el patrimonio y de una sociedad democrática. Sin embargo, en el caso de los archivos de la Web y de redes sociales hay una serie de consideraciones que es necesario anotar cuando nos referimos al acceso y reutilización de grandes volúmenes de datos.

En 2008, Andreas Rauber, Max Kaiser y Bernhard Wachter anotaron que el acceso a las colecciones era aún muy restringido debido a la carencia de herramientas para la interacción y acceso a grandes volúmenes de datos, así como a limitaciones éticas y legales.[285] En poco más de una década, se ha desarrollado *software* para facilitar la gestión y el acceso a los archivos de la Web y a los sets de datos de redes sociales. Inclusive, como se señaló en el capítulo anterior, la acumulación de grandes volúmenes de datos de este tipo de contenidos ha sido utilizada como materia prima para el entrenamiento de herramientas de inteligencia artificial. Las primeras experiencias de uso de los datos masivos de Internet Archive en aplicaciones de Inteligencia Artificial fueron poco consistentes. No obstante, se espera que la evolución de herramientas como IA Copilot de Internet Archive sea una vía para usufructuar los datos guardados en Wayback Machine[286] y establecer criterios éticos para la recuperación de la información.

Por otra parte, el desarrollo de herramientas, cada vez más potentes para el acceso y recuperación de datos de materiales de Archivado Web y redes sociales, enfrenta limitaciones de carácter legal y político abocadas a limitar el acceso abierto a las colecciones digitales.

Una prueba de ello es la demanda legal que interpusieron en 2020, contra el proyecto Open Library de Internet Archive, cuatro editoriales

285 Andreas Rauber, Max Kaiser y Bernhard Wachter, "Ethical issues in Web Archive Creation and Usage – Towards a Research Agenda", en *8th International Web Archiving Workshop (IWAW08)*. 2008, https://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda.

286 Chan, "Navigating Through Archived Websites".

(Hachette, Penguin Random House, Wiley y Harper Collins) en una corte de Nueva York, Estados Unidos, por violar los derechos de autor al ofrecer el acceso y descarga de libros de forma libre y gratuita. En la demanda se arguyó que sin contar con licencia o pago alguno a los autores o editores se escanearon los libros para que pudieran ser descargados con tan solo hacer un *click* en Internet.[287]

Brewster Kahle, fundador de Internet Archive, respondió a la demanda argumentando que Internet Archive funciona como una biblioteca y apoya la publicación, a los autores y lectores.[288]

Las editoriales solicitaron el cierre de Internet Archive porque advirtieron que esta iniciativa les provocó daño económico, aunque no existe evidencia de ello. Hasta marzo de 2023, el juicio continuaba a cargo del juez John G. Koeltl, quien ha rebatido argumentos de ambas partes.[289] El 24 de marzo de 2023, el juez Koeltl falló en favor de las casas publicadoras, rechazando el argumento de uso justo *"fair use"*, presentado por Internet Archive.[290] Al siguiente día se informó que Internet Archive apelaría la decisión.[291]

Si la industria editorial triunfa se reduciría la capacidad del archivo y el acceso público a la información y se favorecería el crecimiento de plataformas privadas. Este juicio expresa una batalla legal y política que utiliza al *copyright* como justificación para defender el uso comercial de la información frente al acceso abierto.

287 Rusell Brandon, "Publishers sue Internet Archive over Open Library ebook lending", The Verge, publicado el 1 de junio de 2020, https://www.theverge.com/2020/6/1/21277036/internet-archive-publishers-lawsuit-open-library-ebook-lending.

288 *Idem.*

289 "Editoriales exigen cierre del Internet Archive en demanda por préstamo de libros digitalizados", R3D Red en Defensa de los Derechos Digitales, publicado el 23 de marzo de 2023, https://r3d.mx/2023/03/23/editoriales-exigen-cierre-del-internet-archive-en-demanda-por-prestamo-de-libros-digitalizados/.

290 Joe Hernandez, "A judge sided with publishers in a lawsuit over the Internet Archive's online library", NPR, publicado el 26 de marzo de 2023, https://www.npr.org/2023/03/26/1166101459/internet-archive-lawsuit-books-library-publishers. El término "fair use" está incluido en la Ley de Derechos de Autor de los Estados Unidos e indica que puede haber instancias en que se puede reproducir parte de una obra por razones que no afectan a quien tiene los derechos.

291 Chris Freeland, "The Fight Continues", *Internet Archive Blogs*, 25 de marzo de 2023, https://blog.archive.org/2023/03/25/the-fight-continues/.

Que unas decisiones fundamentales sobre mantener o destruir los datos estén en manos de actores con fines de lucro, con aspiraciones autocráticas u otros fines tiene enormes consecuencias, no solo para las personas, sino también para la cultura en general. En muchos casos, la pérdida de datos repercute en la producción cultural, la escritura de la historia y, en definitiva, la práctica de la democracia.[292]

Las bibliotecas y archivos desarrollan sus servicios en una tensión permanente. Facilitan el acceso abierto a la información y cuidan de no vulnerar los derechos de autor. Esta tensión con nuevas complejidades se traslada al archivado web y de redes sociales. En razón de lo cual en muchas instituciones de la memoria el acceso a este tipo de materiales se realiza *in situ* y sólo se ofrece la consulta de los metadatos en línea.

Gran parte de los desafíos éticos del archivado web y de redes sociales provienen de los dilemas y discusiones que afronta la Internet. Las colecciones de la Web y de redes sociales constituyen un nuevo tipo documental que no sólo confronta los problemas de las restricciones de derechos de autoría intelectual derivado de su naturaleza, sino que también tiene ante sí aspectos éticos en los que se involucran temas como la privacidad y la protección de datos personales.

Uno de estos ámbitos es el de la privacidad. Muchas personas publican en la Web sin pensar en que esta información sea archivada y pueda ser consultada en el futuro. Es decir, que los autores no están conscientes de que lo publicado puede ser revisado por otra persona en un futuro. El gran volumen de datos que se almacenan derivados del archivado web y de redes sociales requiere del uso de herramientas potentes para la búsqueda y análisis de información. Así uno de los retos más significativos propuestos por Andreas Rauber, Max Kaiser y Berhard Wachter es garantizar que estos archivos no supongan una amenaza para la intimidad de las personas.[293] La privacidad y la protección de los datos personales es un principio ético a considerar para el acceso a los materiales archivados.

292 Rauber, Kaiser y Wachter, "Ethical issues in Web Archive".
293 *Idem*.

Un ejemplo relevante que recupera la disertación antes señalada lo ofrece el Archivo Web de Australia.[294] Establece los siguientes principios éticos que han de tomarse en consideración para restringir el acceso de las páginas web y de redes sociales:

Privacidad de los datos personales

Se ha de procurar que los datos personales sean confidenciales para evitar, en la medida de lo posible, que las personas puedan ser identificadas en situaciones de violencia, robo u otro que la ponga en riesgo.

Datos que vulneren su privacidad

Toda la información que invada la privacidad personal debe ser omitida.

Difamación

La información que agreda o difame a una persona deberá ser restringida.

Derecho comunitario y protocolos culturales

La información sensible que vulnere a determinadas comunidades como son los pueblos originarios o bien las comunidades sexualmente diferenciadas.

Mandato judicial

Cualquier información que por mandato judicial no se puede publicar.

294 "Website category. Restricted content", Trove, https://trove.nla.gov.au/help/categories/websites-category.

Contenido delictivo

Material considerado delictivo en una determinada legislación.

Información dañina

Información que puede ser dañina, aunque no necesariamente sea ilegal. En esta categoría se incluyen imágenes pornográficas, imágenes que promuevan actos terroristas, incitación al odio, etcétera.

Derechos de autor

Contenido sujeto a restricciones derivadas de las limitaciones de derechos de autor.

El respeto a la propiedad intelectual y la protección de datos personales son características que deben ser tomadas en consideración y fundar los servicios para el acceso a este tipo de colecciones. En el caso de las bibliotecas y archivos de alcance nacional se supone, como sucede en el caso de la Biblioteca Nacional de España que los editores y productores de contenido en línea deberán permitir que los centros de conservación recolecten sus publicaciones o facilitar el depósito cuando así se solicite.[295]

Datos gubernamentales protegidos

Algunos materiales están restringidos de forma permanente, mientras que otros se liberarán después de un período de tiempo. Esta recomendación puede ser polémica sobre todo en países donde se limita el acceso a la información con fines políticos.

295 Molina, "Archivo web de las publicaciones", 283-307.

7.3. El bibliotecario y archivista como curadores de contenidos

Frente a un nuevo tipo documental, como el que representan las páginas web ¿en qué institución de la memoria debe recaer la responsabilidad de preservarlas? ¿Cómo llevar a cabo este proceso? ¿Debe ser una tarea que sólo compete a los archivos y bibliotecas?

Las respuestas a las preguntas anteriores son complejas e involucran diversas perspectivas. En esta publicación se pretende orientar a los responsables de archivos y bibliotecas en el uso de directrices para la preservación de páginas web. Y si bien, el presente volumen está orientado a estas instituciones de la memoria, conviene destacar su impronta y utilidad para todos aquellos que producen día a día contenidos digitales en páginas web. En el caso del Archivado Web y de redes sociales, el acceso y la amplia gama de posibilidades de reutilización pueden ser el incentivo y el motivo que sustenta la puesta en marcha de un programa para la selección, acopio, rastreo, conservación, gestión y almacenamiento de este tipo de objetos digitales.

La inserción de innovadores procesos documentales para el tratamiento de documentos, cuyo origen es digital, no solo modifica y ensancha los métodos de trabajo y el perfil profesional, sino añade un rol proactivo en el desarrollo de fondos y colecciones.

Es un lugar común advertir que el archivista, bibliotecario o documentalista se encarga de proteger los documentos desde el momento en que son incorporados para su resguardo y preservación a largo plazo. Esta perspectiva tradicional cambió porque desde hace más de tres décadas se producen, de forma imparable, contenidos nativos digitales que no se compran, donan o entregan en comodato en las instituciones de la memoria. Se publican en Internet y su permanencia es incierta.

La preservación del patrimonio digital contemporáneo requiere de la participación de profesionales de la información que lejos de hacer actividades rutinarias, desarrollen labores proactivas en la creación de

colecciones de valor social, cultural, académico, científico, político, entre otras. Tomando en consideración que con esta tarea se forja un fragmento del patrimonio digital. Condición que conlleva la responsabilidad ética de forjar colecciones con base en principios éticos.

Esto implica la intervención de un curador digital que, en función de una serie de políticas de selección y curaduría, ponga en marcha acciones de acopio y recolección, gestione y almacene datos y ofrezca este material para su acceso y reutilización.

Si bien es cierto que en las dos últimas décadas se han incrementado los estudios e investigaciones en torno al Archivado Web y de redes sociales como un ámbito novedoso en las ciencias sociales; en contraste, se observa que este nuevo perfil profesional no ha permeado como debería en el desarrollo de los bibliotecarios[296] y tampoco de los archivistas. Posiblemente porque estas actividades son nuevas y significan un rompimiento con métodos tradicionales de trabajo.

7.4 Principios deontológicos para el Archivado Web y de redes sociales

Cada profesión se funda en un conjunto de principios y valores comunes que motivan el desempeño cotidiano de sus actividades. Los profesionales de la información que llevan a cabo el archivado de sitios web y de redes sociales, denominados como curadores digitales, comparten valores con otros profesionales que preservan el patrimonio, desarrollan tareas documentales para preservar y dar acceso a la información, así como de aquellos que producen y distribuyen información en medios digitales. Su actuación se enmarca en las diversas recomendaciones formuladas por la Unesco en relación con la protección y acceso universal

296 Francesco D'Amaro, "La memoria digital de España. El Archivo Web como nueva fuente para la historia del presente", en *Del siglo XIX al XXI. Tendencias y debates (Alicante, 20-22 de septiembre de 2018)*, coordinado por Mónica Moreno Seco y editado por Rafael Fernández Sirvent y Rosa Ana Gutiérrez Lloret. Actas del XIV Congreso de la Asociación de Historia Contemporánea (Alicante: Biblioteca Virtual Miguel de Cervantes, 2019), 270-285, https://www.researchgate.net/publication/334319023_La_memoria_digital_de_Espana_El_Archivo_Web_como_nueva_fuente_para_la_historia_del_presente.

al conocimiento. Entre estas se deben citar *Memoria del Mundo. Directrices para la salvaguardia del patrimonio documental;*[297] *Directrices para la preservación del patrimonio digital;*[298] *Directrices Unesco/PERSIST sobre selección del patrimonio digital para su conservación a largo plazo;*[299] *Convención para la Salvaguardia del Patrimonio Cultural Intangible;*[300] *Recomendación sobre la Salvaguardia y la Conservación de las Imágenes en Movimiento*[301] *y la declaración Convirtiendo la amenaza del COVID-19 en una oportunidad para un mayor apoyo al patrimonio documental.*[302]

Comunidades de práctica en el Archivado Web y de redes sociales han abierto espacios de diálogo y debate sobre las implicaciones éticas de dichas prácticas. Cabe destacar el trabajo de Documenting the Now en este tema, cuya misión incluye un compromiso de priorizar prácticas éticas en la preservación de contenidos de redes sociales.[303] En abril de 2018 la organización publicó el texto *Ethical considerations for archiving social media content generated by contemporary social movements: challenges, opportunities, and recommendations.*[304] El documento señala los desafíos principales del archivado de redes sociales, entre los cuales se

297 Unesco, *Memoria del Mundo. Directrices para la salvaguardia del patrimonio documental*, edición revisada y preparada por Ray Edmondson (París: Unesco, 2002), 72, https://unesdoc.unesco.org/ark:/48223/pf0000125637_spa.

298 Unesco, *Directrices para la preservación del patrimonio digital*, preparado por la Biblioteca Nacional de Australia, marzo 2003, 170, https://unesdoc.unesco.org/ark:/48223/pf0000130071_spa.

299 Unesco, *Directrices Unesco/PERSIST sobre selección del patrimonio digital*.

300 Unesco, "Convención para la Salvaguardia del Patrimonio Cultural Intangible", 32 Conferencia General de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, celebrada en París, del 29-septiembre al 16-octubre de 2003, Unesco Patrimonio Cultural Inmaterial, https://ich.unesco.org/es/convenci%C3%B3n.

301 Unesco, *Recomendación sobre la Salvaguardia y la Conservación de las Imágenes en Movimiento*, 21 Conferencia General de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, celebrada en Belgrado, del 23-septiembre al 28-octubre de 1980, Unesco, https://www.unesco.org/es/legal-affairs/recommendation-safeguarding-and-preservation-moving-images.

302 Unesco. *Convirtiendo la amenaza del COVID-19 en una oportunidad para un mayor apoyo al patrimonio documental*, Unesco, publicado el 5 de abril de 2020, https://www.unesco.org/en/articles/turning-threat-covid-19-opportunity-greater-support-documentary-heritage.

303 "Documenting the Now", DN.

304 Bergis Jules, Ed Summers y Vernon Mitchell Jr., "Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities and Recommendations" (Documenting the Now White Paper), 1-12, Documenting the Now, publicado en abril de 2018, https://www.docnow.io/docs/docnow-whitepaper-2018.pdf.

encuentran el desconocimiento de los usuarios sobre la posibilidad de que sus contenidos sean preservados y la posibilidad de uso fraudulento de los contenidos.[305] Una de las recomendaciones principales se centra en priorizar la colaboración con creadores de contenido y con las comunidades que desean documentar.[306]

La política de Documenting the Now en cuanto al acceso a sets de datos de tuits busca lograr el complicado balance entre el acceso a la información y el respeto a la privacidad de los creadores de contenido. En su catálogo de tuits,[307] Documenting the Now solo da acceso a los identificadores de los tuits acopiados como una manera de cumplir los términos de uso de Twitter, el cual prohíbe hacer público los datos recopilados, pero, además, para respetar el derecho de los usuarios de eliminar sus tuits o hacerlos privados. Esto porque de las listas de identificadores de tuits, el usuario solo puede recuperar tuits que se mantengan públicos. Como expresa Documenting the Now:

> Nuestro desarrollo tecnológico se ha centrado en los derechos de los investigadores de compartir sus colecciones para re-uso, mientras que también respeta los derechos de los creadores de contenido de proteger o eliminar su contenido de la vista pública. El proceso de publicar los identificadores de tuits, validado por Twitter, ha sido piedra angular para buscar atender ambas metas.[308]

La colaboración con creadores de contenidos es una ruta importante para el archivado de web y de redes sociales, y un aspecto que va generando más relevancia ante las decisiones de las empresas de redes sociales de restringir el acceso a sus API.

La elaboración de los principios deontológicos para el archivado web y de redes sociales deberá ser una construcción colectiva que derive de la observación y análisis de los procedimientos y del contexto en que se

305 *Ibid*., 3.
306 *Ibid.,* 12.
307 "Catalog", DN.
308 Jules, Summers y Mitchell Jr., "Ethical Considerations for Archiving", 11.

desarrolla esta actividad profesional. En consecuencia, conviene señalar que los documentos de la Web y los datos de las redes sociales son testimonios de la sociedad y deben ser preservados con la misma importancia que tienen otro tipo de documentos. Su riesgo de desaparición es alto y por ello su salvaguarda debe ser una tarea inminente. Para cumplir ese cometido es necesario que se dote de la formación, recursos tecnológicos y financieros necesarios a los responsables de llevar a cabo esta tarea.

Los curadores digitales debieran ser los responsables de garantizar la integridad, autenticidad, accesibilidad y uso de los documentos de la Web, así como de los sets de datos de las redes sociales, resguardándolos de alteraciones deliberadas, censura, desastres naturales e indiferencia institucional.

De forma análoga a lo señalado por Ray Edmondson en relación con los archivos sonoros y audiovisuales, en lo que concierne a los documentos de la Web y de redes sociales:

> Su selección, su protección y su accesibilidad en nombre del interés público han de regirse por normas objetivas y no por presiones políticas, económicas, sociológicas o ideológicas, como, por ejemplo, el concepto de corrección política que está en vigor. El pasado ha quedado fijado. No puede cambiarse.[309]

Aunque esta tarea es un complejo desafío para los curadores digitales en el contexto actual de posverdad.

La posverdad ocasiona que la falsedad se confunda con la realidad. Así las mentiras repetidas de forma sistemática se convierten en realidad, en tanto que la verdad se diluye ante el exceso de información. El problema es de tal magnitud que en un estudio realizado entre 2006 y 2007, en relación con la información falsa distribuida a través de Twitter se encontró que las noticias falsas tienen un 70 % más de probabilidades de ser retuiteadas que las verdaderas.[310] En múltiples investigaciones se

309 Edmondson, *Filosofía y principios,* 10.

310 Soroush Vosoughi, Deb Roy y Sinan Aral, "The Spread of True and False News Online", MIT Initiative on the Digital Economy, publicado en marzo de 2018, (sección Publications/Research Brief), https://ide.mit.edu/wp-content/uploads/2018/12/2017-IDE-Research-Brief-False-News.pdf.

ha advertido las afectaciones que la posverdad ocasiona en las democracias. La democracia requiere al menos de información y de un discurso público verídico,[311] principio que se ve amenazado por los medios de comunicación y especialmente por las redes sociales. Los políticos y empresarios aprovechan la desinformación a su favor. Lo que implica que el Archivado Web y de redes sociales no es una tarea documental sistemática y rutinaria, sino que demanda tener una perspectiva crítica e informada.

La perspectiva del curador digital es crucial para identificar y diferenciar la información veraz, dado que la información que se capture será la que narra y testifica la historia contemporánea. Esta situación fue advertida hace tiempo y por ello, se han comenzado a desarrollar herramientas para analizar las fuentes de información. Un ejemplo es *Twittervane*, herramienta capaz de analizar las fuentes de Twitter y determinar qué sitios web se comparten con mayor frecuencia en función de un tema determinado durante un tiempo específico. Estos sitios web[312] se pueden proporcionar a los curadores para facilitar el proceso de selección.[313]

La selección de la información de la Web y de redes sociales que se va a preservar se habrá de distinguir entre lo que es falso y lo que es real.[314] Con base en esos datos se estudiará, investigará y nos conocerán las generaciones por venir.

Es así que el trabajo curatorial no está exento de tomar en consideración en su rutina diaria, la tensión entre las fuerzas que siempre desean reescribir o censurar la historia para proteger sus intereses particulares.[315]

Por otra parte, los curadores digitales también confrontan la tensión entre la propiedad intelectual de los contenidos publicados en la Web y

311 Rafael Benítez, "Michael Sandel: El peligro no es que sea difícil distinguir lo real de lo falso, sino que esa distinción deje de importarnos", *Telos,* n.° 122 (junio de 2023), https://telos.fundaciontelefonica.com/wp-content/uploads/2023/06/telos-122-entrevista-posverdad-michael-sandel.pdf.

312 Helen Hockx-Yu, "Evaluating Twittervane", IIPC, https://netpreserve.org/projects/evaluating-twittervane/.

313 "Twittervane", IIPC/Twittervane, https://github.com/iipc/twittervane.

314 Benítez, "Michael Sandel: El peligro", *Telos*.

315 Edmondson, *Filosofía y principios,* 11.

en redes sociales y su acceso abierto. Deben considerar esta dicotomía para no vulnerar los derechos de autor, morales y patrimoniales que, por la naturaleza hipermedia de este tipo de documentos, involucran a diversos autores.

Los modos de acceso y consulta de la Web y redes sociales son diferentes a los que han sido instalados en bibliotecas y archivos. Derivado de lo cual, en muchos casos los programas de archivado web y de redes sociales se circunscriben a su consulta dentro de las bibliotecas y archivos con fines de investigación y docencia. En consecuencia, los curadores digitales salvan obstáculos para preservar y dar acceso a información veraz, cuyos usos benefician a la sociedad y fortalecen las democracias.

La salvaguarda del patrimonio digital está respaldada por el Depósito Legal que ha ampliado su ámbito de responsabilidad de los libros, a los soportes sonoros y audiovisuales y de forma reciente a los documentos de origen digital como son los sitios web.

La curaduría digital es una profesión en formación que se funda en la combinación de saberes. Se sostiene en la vasta experiencia y conocimiento adquirido en bibliotecas, archivos y museos. Y a la vez, involucra y requiere del uso de tecnologías digitales durante todo el ciclo de vida. Es necesario señalar que recupera los saberes recientes derivados de la preservación digital de sendas colecciones sonoras y audiovisuales. Los primeros esfuerzos fueron impulsados por bibliotecas nacionales y los desarrolladores de la tecnología. Gracias a ello, en la actualidad contamos con tecnología madura para el acopio, rastreo, almacenamiento y acceso. La evolución de las herramientas ha buscado calidad, sencillez y eficiencia en estas tareas documentales.

Este tipo de archivado se desarrolla en cada institución de forma diferente, en función de cómo se incorpora esta actividad en las organizaciones. Por ejemplo, en la Biblioteca Nacional de Francia, el equipo de archivado web se dividió entre los departamentos de Depósito Legal y Tecnologías de la Información. En tanto, en la Biblioteca Británica, en un principio esta tarea se llevó a cabo entre el área de tecnología y

las secciones de las colecciones; más adelante se reagrupó en una sola sección. Algunas instituciones deciden recurrir a contratar a terceros, es decir, empresas especializadas para realizar cosechas masivas. Un ejemplo de ello, es la Deutsche Nationalbibliothek (Biblioteca Nacional de Alemania) que trabaja con la empresa OIA.[316]

En 2016 la National Digital Stewardship Alliance (NDSA) de Estados Unidos presentó los resultados de una encuesta realizada en 2016 en torno al personal destinado para llevar a cabo el archivado web. Se observó que esta tarea se incluye como parte de otras actividades que se desarrollan en las bibliotecas y solo se destina un cuarto del tiempo dedicado a actividades de preservación.[317] Esto significa que las prácticas relativas a la salvaguarda de información de origen digital aún no están incorporadas en las prácticas de preservación. Cada institución de acuerdo con sus recursos humanos y financieros destina personal para desarrollar esta tarea.

El Archivado Web y de redes sociales podría ser visto como una prolongación de las actividades de los bibliotecarios, archivistas y documentalistas. Los sitios web pueden ser considerados como una extensión que complementa a las colecciones físicas. En todo caso, son materiales que se deben preservar indudablemente.[318]

Entre las habilidades y competencias más importantes que debe tener el personal que realiza actividades de archivado web y de redes sociales, destacan el manejo de herramientas para el acopio, gestión, calidad y acceso. Es decir, la configuración del *software* o *crawler* para el rastreo o recogida de información; así como, el manejo de archivos en formato WARC y su incorporación a la plataforma de gestión y acceso. Además de las habilidades necesarias para la curaduría, evaluación y selección, también es importante considerar la verificación de la calidad de la información.[319]

316 Peter Stirling y Jaanus Kõuts, "How to fit in? Integrating a web archiving program in your organization", Workshop report and evaluation, Framework Education and Training BnF, Paris, France, November 26-30 2012, IIPC, publicado en abril de 2013, https://netpreserve.org/resources/IIPC_project-Report_and_Evaluation_of_BnF_IIPC_Workshop_on_Web_Archiving.pdf.

317 Bailey *et al.*, *Web Archiving in the United States: A 2016 Survey*.

318 Stirling y Kõuts, "How to fit in?".

319 Bailey *et al.*, *Web Archiving in the United States: A 2016 Survey*.

Las actividades intelectuales y el manejo de tecnología para el manejo y gestión de documentos de origen digital son habilidades que deben sumarse en la formación de bibliotecarios, archivistas y documentalistas para la creación de colecciones digitales. Estas destrezas se expanden más allá del ámbito de las disciplinas abocadas a la preservación del patrimonio. Por ello, investigadores de disciplinas como la historia, la antropología y la sociología podrían incorporar su experiencia en la creación y estudio de colecciones digitales,[320] así como de humanidades digitales. Se anota entonces un nuevo perfil profesional en desarrollo: el curador de contenidos digitales.

El *content curator* (curador de contenidos) se refiere a la actividad de selección y difusión de contenidos de la Web. Sin la existencia de la Web y de las redes sociales no se podría hablar de la curaduría de contenidos.

Desde hace más de una década se ha documentado la participación de bibliotecarios y archivistas en tareas de recolección y acopio de páginas web.[321] En el caso de las bibliotecas, la producción de materiales analógicos ha decaído y el Archivado Web constituye una oportunidad para que los archivistas y bibliotecarios ensanchen sus fronteras de trabajo documental.

Hay un desigual trabajo de archivado web y de redes sociales en el mundo y por ello, también los programas de formación y actualización profesional son limitados en países donde apenas se sabe de esta forma documental. Esto incide en que la comunidad de curadores, archivistas o profesionales de la información que desarrollan esta tarea sea pequeña. Y a pesar de que se trata de una formación reciente se ha fortalecido porque ha impulsado muchos proyectos de colaboración y cooperación internacional.

Los esfuerzos aislados emprendidos por las bibliotecas nacionales se han consolidado con la fundación del International Internet Preservation Consortium (IIPC), creado en 2003 de forma análoga a otras grandes

320 D'Amaro, "La memoria digital de España", 270-285.

321 Hanna, "El Modelo".

organizaciones internacionales de bibliotecarios y archivistas como la Federación Internacional de Bibliotecas y Bibliotecarios (IFLA), la Asociación Internacional de Archivos Sonoros y Audiovisuales (IASA), la Federación Internacional de Archivos de Televisión (FIAT) y la Federación Internacional de Archivos Fílmicos (FIAF). Este tipo de organizaciones son fundamentales porque se comparten problemas comunes, se diseñan proyectos e iniciativas de colaboración.

Con base en lo anterior, se sugiere que el curador digital incorpore en su práctica profesional entre otros valores: la veracidad en el tratamiento de la información, la honestidad en la selección de la información, imparcialidad y precisión en el establecimiento de los parámetros para el acopio y recolección; transparencia en el desarrollo de las actividades de archivado web y redes sociales; disposición de colaboración y cooperación con otros colegas del gremio para mejorar procesos; rendición de cuentas del trabajo documental realizado, problemas y soluciones, respeto a la confidencialidad y privacidad de los datos personales archivados.

Conclusiones

La preservación no es solamente un compromiso de conservación del pasado. Es el medio para comprender el presente y forjar el futuro. En la segunda década del nuevo milenio, la historia contemporánea debe ser documentada día a día y para ello, es necesario establecer metodologías y emplear herramientas para el acopio, conservación y acceso a contenidos de origen digital como los que se publican en Internet.

Este contexto contemporáneo conmina a los profesionales de la información a pensar en torno a sus roles como agentes activos en procesos de documentación del presente y en la necesidad de diseñar formas de colaboración con otros grupos o instituciones que coincidan en el interés por salvaguardar el patrimonio digital.

El Archivado Web y de redes sociales es el modo de documentación hipermedia de materiales de origen digital como son las páginas web y las redes sociales. Los primeros esfuerzos de Archivado Web se sitúan en la última década del siglo pasado a nivel internacional, pero su incorporación como ámbito de trabajo en las instituciones de la memoria ha sido desigual en el mundo. En algunos países, desde hace más de tres décadas, se emprendieron iniciativas para salvaguardar la Web, pero en otros aún es una práctica desconocida o poco implementada.

En México el Archivado Web y de redes sociales es una práctica documental inexplorada y su abordaje como tema de investigación ha sido poco estudiado. Esta condición conlleva a que se pierdan de forma sistemática sendas colecciones de páginas web e información de redes sociales que dan cuenta de la historia reciente.

El Archivado Web y de redes sociales es una práctica que no se puede desestimar, debe formar parte activa de las tareas documentales ya que es un tema de relevancia social por la diversidad y volumen de contenidos digitales que día a día se publican; la permanencia de estos materiales a largo plazo es un desafío porque su preservación no es una práctica cotidiana entre las instituciones de la memoria que llevan a cabo este tipo de acciones.

Por ello, este libro no es una obra concluyente, sino que pretende generar nuevas vías de investigación, atendiendo los cambios constantes en las maneras en que se producen, transmiten y controlan los contenidos en la Web y en las redes sociales. Por lo tanto, existe una amplia gama de posibilidades de abrir estas nuevas vías de investigación por medio de proyectos colaborativos e interdisciplinarios, tanto con instituciones mexicanas como de otros países latinoamericanos.

Nuestra experiencia en el proceso de acopio y curaduría de contenidos de Twitter (ahora X) es un ejemplo de estos cambios constantes, donde las decisiones de la empresa luego de ser adquirida por Elon Musk han comenzado a impactar el acceso a la API de la red social, lo que a su vez afecta las metodologías de acopio empleadas (ver capítulo V). Ante estos cambios, comunidades de práctica, como Documenting the Now han estado reflexionando sobre las implicaciones de estas decisiones, pero han sido conversaciones provenientes principalmente de Estados Unidos.

Por otra parte, es indudable que la preservación digital de la Web y de redes sociales tiene implicaciones en el medioambiente porque es necesario asegurar el almacenamiento digital seguro de grandes volúmenes de datos que requieren de energía ininterrumpida y de un cambio

sistemático de infraestructura tecnológica, para intentar afrontar a tiempo la obsolescencia tecnológica. Y, además, porque cada vez que una persona realiza una consulta en línea se afecta al medioambiente y se deja una huella de carbono.

La preservación digital de la Web y de redes sociales, así como de otro tipo de soportes, es un factor que afecta al cambio climático. Por ello, la búsqueda de soluciones debe ser una de las prioridades de las instituciones de la memoria que preocupadas por la salvaguarda de este patrimonio deben a la vez sugerir alternativas ecológicas para intentar revertir, en la medida de lo posible, este daño. El tema supera los alcances de la publicación que el lector tiene en sus manos, pues amerita un estudio a profundidad por su relevancia contemporánea.

El presente volumen fue escrito para ser consultado por bibliotecólogos, archivistas y profesionales de la información que están conscientes de la importante transformación que en cuanto al tratamiento documental se desarrolla en las instituciones de la memoria y que vislumbran nuevos procesos, flujos de trabajo y roles profesionales como el curador de contenidos digitales.

Sería recomendable que los procesos, técnicas y tecnologías empleados en el Archivado Web y de redes sociales se incorporen en la currícula de las universidades que forman a los profesionales de la información en el tratamiento de la información digital. Como parte de este proceso de formación, el estudiantado debe exponerse además a espacios que permitan dar una mirada crítica sobre el Archivado Web y sus implicaciones teórico-metodológicas dentro de las Ciencias de la Información. Los programas académicos juegan además un papel fundamental para la capacitación y actualización de profesionales de la información por medio de talleres y diplomados enfocados en las prácticas de archivado web.

Es precisamente sobre estas últimas que se han de tomar en consideración los derechos de autor de los contenidos recopilados. Este es sin lugar a dudas, uno de los temas que deberán ser revisados con mayor

atención por quienes decidan iniciar actividades de Archivado Web y redes sociales. Otro tema fundamental a ser estudiado a mayor profundidad es el de la selección en el contexto de proyectos de Archivado Web, evaluando cómo estrategias de valoración, desde la bibliotecología y la archivística, podrían ser aplicadas al archivado web. Aun cuando el contexto complejo y cambiante de la Web y de las redes sociales abre las puertas a repensar conceptos y prácticas bibliotecológicas y archivísticas, estos mismos fundamentos pueden presentar pistas sobre cómo documentar el presente. Por ejemplo, la implementación de prácticas de adquisición por medio de donaciones, donde se establece una relación creador-institución, presenta una potencial estrategia para atender temas de acopio, derechos de autor y dependencia a las corporaciones que controlan las redes sociales.

El acceso a los contenidos recopilados como resultado de prácticas de Archivado Web y de redes sociales posee uso potencial para la generación de nuevas investigaciones y para la docencia. Por lo tanto, esta investigación demostró la necesidad de incorporar como nuevas prácticas de documentación en bibliotecas y archivos mexicanos la preservación de páginas web y de redes sociales. Así como impulsar el debate y la inserción del Archivado Web y de redes sociales en la Ley de Depósito Legal nacional.

Bibliografía

Acker, Amelia, y Adam Kriesberg. "Tweets may be archived: Civic engagement, digital preservation and Obama White House social media data". *Proceedings of the Association for Information Science and Technology* 54, n.° 1 (octubre de 2017): 1–9. DOI: 10.1002/pra2.2017.14505401001.

_____. "Social media data archives in an API-driven world". *Archival Science* 20, n.° 2 (2020): 105–123. DOI: 10.1007/s10502-019-09325-9.

Adamo Idoeta, Paula. "Por qué los algoritmos de las redes sociales son cada vez más peligrosos". *BBC News Mundo*, 12 de octubre de 2021. https://www.bbc.com/mundo/noticias-58874170.

Adeyemo, Adeola. "History & Importance of the First Social Media Site - Six Degrees". Adeola. Publicado en 2018. https://adeolawrites.com/history-importance-of-the-first-social-media-site-six-degrees/.

Aguilera, Miguel de, y Andreu Casero-Ripollés. "¿Tecnologías para la transformación? Los medios sociales ante el cambio político y social. Presentación". *Icono 14. Revista de Comunicación y Tecnologías Emergentes* 16, n.º 1 (2018): 1-21. https://doi.org/10.7195/ri14.v16i1.1162.

Amnistía Internacional España. "Amnistía Internacional publica un nuevo informe sobre Ferguson que documenta abusos contra los derechos humanos". Publicado el 24 de octubre de 2014. https://www.es.amnesty.org/en-que-estamos/noticias/noticia/articulo/amnistia-internacional-publica-un-nuevo-informe-sobre-ferguson-que-documenta-abusos-contra-los-derec/.

Archipelago Documentation. "Archipelago Commons Intro". Última modificación el 18 de octubre de 2021. https://docs.archipelago.nyc/1.0.0/.

Archive-It. Consultado el 9 de junio de 2023. https://archive-it.org/.

_____. "About Archive-It". Consultado el 30 de marzo de 2023. https://archive-it.org/learn-more/.

_____. "Archive-It Sponsored". Consultado el 7 de marzo de 2023. https://archive-it.org/blog/archive-it-sponsored/.

_____. "Spontaneous Event Collections". Consultado el 8 de marzo de 2023. https://archive-it.org/blog/spontaneous-events/.

ArchiveTeam/grab-site. "Grab-site". Consultado el 10 de junio de 2023. https://github.com/ArchiveTeam/grab-site.

Ariel, Yaron y Ruth Avidar. "Information, Interactivity and Social Media". *Atlantic Journal of Communication* 23, n.º 1 (2015): 19-30. https://doi.org/10.1080/15456870.2015.972404.

Arvidson, Allan, Krister Persson y Johan Mannerheim. "The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of 'complete' collection of web pages". 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 de agosto de 2000. https://archive.ifla.org/IV/ifla66/papers/154-157e.htm.

AT Archiveteam. "GeoCities". Última modificación el 20 de enero de 2023. https://wiki.archiveteam.org/index.php/GeoCities#:~:text=GeoCities%20was%20a%20once%20very,on%20the%20World%20Wide%20Web.

Bailey, Jefferson, Abigail Grotke, Edward McCain, Christie Moffatt y Nicholas Taylor. *Web Archiving in the United States: A 2016 Survey.* An NDSA Report. National Digital Stewardship Alliance. Results of a Survey of Organizations Preserving Web Content February 2017. https://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf.

Bailey, Jefferson, director de Servicios de Datos de Internet Archive. Entrevistado por Joel Antonio Blanco-Rivera, el 14 de julio de 2023.

BBC News Mundo. "MySpace: el error en un servidor que hizo perder a la red social 12 años de música almacenada", 18 de marzo de 2019. https://www.bbc.com/mundo/noticias-47612172.

_____. "Six Degrees: cómo fue y quién creó la primera red social de internet, inspirada por la teoría de los «seis grados»", 8 de junio de 2019. https://www.bbc.com/mundo/noticias-48558989.

Benítez, Rafael. "Michael Sandel: El peligro no es que sea difícil distinguir lo real de lo falso, sino que esa distinción deje de importarnos". *Telos*, n.º 122 (junio de 2023): 26-34. https://telos.fundaciontelefonica.com/wp-content/uploads/2023/06/telos-122-entrevista-posverdad-michael-sandel.pdf.

Berlin, John A., Mat Kelly, Michael L. Nelson y Michele C. Weigle. "WAIL: Collection-Based Personal Web Archiving". En *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada*, editado por IEEE, 2017, 1-2. DOI: 10.1109/JCDL.2017.7991619.

Berners-Lee, Tim, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen y Arthur Secret. "The World-Wide Web". *Communications of the ACM* 37, n.º 8 (agosto de 1994): 76–82. https://doi.org/10.1145/179606.179671.

Biblioteca Nacional de Dinamarca. "Netarkivet". Det Kgl. Bibliotek. https://www.kb.dk/en/find-materials/collections/netarkivet.

Biblioteca Nacional de España. "Archivo de la Web española". BNE. https://www.bne.es/es/colecciones/archivo-web-espanola.

_____. "Historia de la colección". BNE https://www.bne.es/es/colecciones/archivo-web-espanola.

Biblioteca Nacional de Francia. "Consulter les Archives de l'internet". BnF. https://www.bnf.fr/fr/archives-de-linternet.

Blanco-Rivera, Joel Antonio. "Curaduría digital y la preservación de contenidos web: creando una colección de tuits sobre la huelga de la Universidad de Puerto Rico". Encuentro Latinoamericano de Bibliotecarios, Archivistas y Museólogos. Revalorizando el patrimonio en la era digital, 9-13 de octubre de 2017. https://www.institutomora.edu.mx/EBAM/2017/Ponencias/Curaduria%20digital%20y%20la%20preservacion%20de%20contenidos%20web%20creando%20una%20coleccion%20de%20tuits%20sobre.pdf.

_____. "La archivología en el contexto de la sociedad interconectada por redes". *Revista Interamericana de Bibliotecología* 42, n.º 3 (2019): 213-221. https://doi.org/10.17533/udea.rib.v42n3a02.

Blanco-Rivera, Joel Antonio, Irmarie Fraticelli Rodríguez y Marisol Ramos. "Documentando lo espontáneo: las protestas #RickyRenuncia". *Archidata: Boletín de la Red de Archivos de Puerto Rico* 18, n.º 1 (octubre de 2020): 13-17. https://archiredpr.files.wordpress.com/2020/11/archidata2020.pdf.

Boyd, Danah M. y Nicole B. Ellison. "Social Network Sites: Definition, History and Scholarship". *Journal of Computer-Mediated Communication* 13, n.º 1 (1 de octubre de 2007): 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Bragg, Molly y Kristine Hanna. "The web archiving life cycle model". The Archive-It Team Internet Archive. Publicado en marzo de 2013. https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf y https://archive.org/details/WALCM.

Brandon, Rusell. "Publishers sue Internet Archive over Open Library ebook lending". The Verge. Publicado el 1 de junio de 2020. https://www.theverge.com/2020/6/1/21277036/internet-archive-publishers-lawsuit-open-library-ebook-lending.

Brown, Adrian. *Archiving Websites: A Practical Guide for Information Management Professionals.* Reino Unido: Facet Publishing, 2006.

Browsertrix Cloud. "Automated Browser-Based Crawling at Scale". https://browsertrix.cloud/features/.

Bruns, Axel y Stefan Stieglitz. "Twitter data: what do they represent?". *It-Information Technology* 56, n.º 5 (2014): 240–245. DOI: 10.1515/itit-2014-1049.

Bruns, Axel y Katrin Weller. "Twitter as a first draft of the present: and the challenges of preserving it for the future". En *WebSci '16: Proceedings of the 8th ACM Conference on Web Science*, 183–189. NY, USA: Association for Computing Machinery, 2016. https://doi.org/10.1145/2908131.2908174.

Carr, Caleb T. y Rebecca A. Hayes. "Social media: Defining, Developing and Divining". *Atlantic Journal of Communication* 23, n.° 1 (2015): 46-65. DOI: 10.1080/15456870.2015.972282.

Caswell, Michelle, Marika Cifor y Mario H. Ramirez. "«To Suddenly Discover Yourself Existing»: Uncovering the Impact of Community Archives". *The American Archivist* 79, n.° 1 (2016): 56-81. DOI: 10.17723/0360-9081.79.1.56.

Chan, Peter. "Navigating Through Archived Websites: From Text Matching to Generative AI-Enhanced Q&A". International Internet Preservation Consortium. Publicado el 28 de junio de 2023. https://netpreserveblog.wordpress.com/2023/06/28/navigating-through-archived-websites-from-text-matching-to-generative-ai-enhanced-qa/.

Cherre, Ilan K. "The Internet Archive está colapsado y la culpa la tiene una IA en proceso de entrenamiento". Computer Hoy. Publicado el 29 de mayo de 2023. https://computerhoy.com/internet/internet-archive-colapsado-culpa-tiene-ia-proceso-entrenamiento-1252426.

Cook, Terry. "Mente sobre la materia: hacia una nueva teoría de la valoración archivística". *Revista d'Arxius,* n.° 3 (2004): 119-154. http://arxiversvalencians.org/wp-content/uploads/2020/04/revista2004_cook.pdf.

Cooper, Belle Beth. "The Surprising History of Twitter's Hashtag Origin and 4 Ways to Get the Most out of Them". Buffer. Publicado el 24 de septiembre de 2013. https://buffer.com/resources/a-concise-history-of-twitter-hashtags-and-how-you-should-use-them-properly/.

Costa, Miguel, Daniel Gomes y Mário J. Silva. "La evolución del archivo web". *Revista Internacional de Bibliotecas Digitales* 18, (2017): 191–205. https://doi.org/10.1007/s00799-016-0171-9.

D'Amaro, Francesco. "La memoria digital de España. El archivo web como nueva fuente para la historia del presente". En *Del siglo XIX al XXI. Tendencias y debates (Alicante, 20-22 de septiembre de 2018),* coordinado por Mónica Moreno Seco y editado por Rafael Fernández Sirvent y Rosa Ana Gutiérrez Lloret, 270-285. Actas del XIV Congreso de la Asociación de Historia Contemporánea. Alicante: Biblioteca Virtual Miguel de Cervantes, 2019.

Da Silva, Chantal. "Twitter rebrands to 'X' as Elon Musk Loses iconic bird logo". NBC News, 24 de julio de 2023. https://www.nbcnews.com/news/us-news/twitter-rebrands-x-elon-musk-loses-iconic-bird-logo-rcna95880#.

Day, Michael. "The Long-Term Preservation of Web Content". En *Web Archiving*, editado por Julien Masanès, 177-194. Berlin: Springer, 2006.

DCC Because good research needs good data. "What is digital curation?". Consultado el 25 de mayo de 2023. https://www.dcc.ac.uk/about/digital-curation.

DC Public Library. "Dig DC API Documentation". https://dcpubliclibrary.github.io/digdc/.

Developer Portal. "Twitter's v2 API". https://developer.twitter.com/en/portal/products/free.

DHNB Digital Humanities in the Nordic and Baltic Countries. "Kulturarw3 – The Web Archive of the National Library of Sweden". Consultado el 18 de mayo de 2023. https://dhnb.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden/.

DIGDC. "COVID-19 in Washington, D.C. Twitter Archive". https://digdc.dclibrary.org/islandora/object/dcplislandora%3A237558?solr_nav%5Bid%5D=7bf56a3c0a50a9608922&solr_nav%5Bpage%5D=0&solr_nav%5Boffset%5D=1.

Digital Curation Centre. "What is digital curation?". Consultado el 25 de mayo de 2023. https://www.dcc.ac.uk/about/digital-curation.

Digital Preservation Coalition. "Storage". DPC. https://www.dpconline.org/handbook/organisational-activities/storage.

Dixon, Stacy Jo. "Facebook: quarterly number of MAU (monthly active users) worldwide 2008-2023". Statista. Publicado el 9 de mayo de 2023. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

DN. "Catalog". https://catalog.docnow.io.

_____. "Documenting the Now". https://www.docnow.io.

DN twarc. "Twarc". https://twarc-project.readthedocs.io/en/latest/.

DocumentingTheNow [@documentnow]. "The Hydrator app has recently had its keys revoked. The extremely limited read access in Twitter's new API quotas mean it will no longer work. Hydrator has operated for the last 7 years to help users reconstitute tweet ID datasets, but no more...". Tweet, publicado el 18 de abril de 2023. https://twitter.com/documentnow/status/1648325488236961792?s=20.

Donovan, Lori. "Japan Disaster Archives: Collaboration for successful web archiving". Archive-It. Publicado el 28 de febrero de 2013. https://archive-it.org/blog/post/japan-disaster-archives-collaboration-for-successful-web-archiving/.

Edmondson, Ray. *Filosofía y principios de los archivos audiovisuales,* 1.ª ed. en español. Unesco, IIBI-UNAM, UASLP, 2018. https://www.mowcapunesco.org/wp-content/uploads/Philos-3-Spanish-2018.pdf.

EFE. "Sentir miedo al volver a casa por las noches…". *Heraldo,* 29 de abril de 2018. https://www.heraldo.es/noticias/nacional/2018/04/28/miles-mujeres-relatan-sus-agresiones-sexuales-twitter-animadas-bajo-lema-cuentalo-1240464-305.html.

El Comercio. "La Manada: ¿cómo acabó el repudiable caso de violación grupal que conmocionó a España?", 20 de octubre de 2020. https://elcomercio.pe/mundo/europa/la-manada-como-acabo-el-repudiable-caso-de-violacion-grupal-que-conmociono-a-espana-noticia/.

EP. "El asesinato de Michael Brown y los disturbios raciales en Ferguson: todas las claves". *20minutos,* 19 de agosto de 2014. https://www.20minutos.es/noticia/2217996/0/claves-asesinato-michael-brown/disturbios-raciales-eeuu/ferguson-misuri/.

Esmero/Archipelago-Documentation. "Archipelago Commons Documentation Repository". https://github.com/esmero/archipelago-documentation.

European Commission. "Networked European deposit library". Consultado el 18 de mayo de 2023. https://cordis.europa.eu/project/id/LB5648.

Fallarás, Cristina, Aniol Maria, Vicenç Ruiz, Karma Peiró, Fernando Cucchietti y BSC (Barcelona Supercomputing Center). El Proyecto #Cuéntalo. https://www.bsc.es/viz/cuentalo/.

Farrell, Matthew, Edward McCain, Maria Praetzellis, Grace Thomas y Paige Walker. *Web Archiving in the United States: A 2017 Survey.* An NDSA Report. National Digital Stewardship Alliance. Results of a Survey of Organizations Preserving Web Content Octubre 2018. https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/17590/Web%20Archiving%20in%20the%20United%20States_A%202017%20Survey.pdf.

Foscarini, Fiorella. "Archival Appraisal in Four Paradigms". En *Currents of Archival Thinking*. 2.ª ed. Editado por Heather MacNeil y Terry Eastwood, 107-133. Libraries Unlimited, ABC-CLIO, 2017. https://publisher.abc-clio.com/9781440839092/5.

Freeland, Chris. "The Fight Continues". *Internet Archive Blogs,* 25 de marzo de 2023. https://blog.archive.org/2023/03/25/the-fight-continues/.

García-Marín, David y Marta Merino-Ortego. "Desinformación anticientífica sobre la COVID-19 difundida en Twitter en Hispanoamérica". *Cuadernos.info,* n.º 52 (mayo 2022): 24-46. https://doi.org/10.7764/cdi.52.42795.

Gilliland, Anne J. "Reconceptualizing Records, the Archive and Archival Roles and Requirements in a Networked Society". *Knygotyra* 63 (january 2014): 17-34. DOI: 10.15388/kn.v63i0.4011. https://www.journals.vu.lt/knygotyra/article/view/4011/2773.

GitHub. "DocNow/twarc". Consultado el 23 de marzo de 2023. https://github.com/DocNow/twarc/blob/main/utils/media2warc.py.

GNU Operating System. "GNU Wget". Consultado el 9 de junio de 2023. https://www.gnu.org/software/wget/.

Goel, Vinay. "Defining Web pages, Web sites and Web captures". *Internet Archive Blogs,* 23 de octubre de 2016. "https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.

González Flórez, John Alexander. *Mini/Manual. Archivamiento web: conceptos básicos, estrategias y mejores prácticas.* Bogotá: Archivo General de la Nación Colombia, 2015. https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicacionees/ArchivamientoWeb.pdf.

Goos, Arnoud. "Archiving broadcasters' websites a discussion of web archiving as context to the radio and television collection". En *2015 Digital Heritage, Granada, Spain.* IEEE Xplore, (2015): 627-630. DOI: 10.1109/DigitalHeritage.2015.7419584.

Guallar, Javier y Javier Leiva-Aguilera. *El content curator. Guía básica para el nuevo profesional de Internet.* Barcelona: Universitat Oberta de Catalunya, 2013. Colección El profesional de la información 24.

Hanna, Kristine. "El Modelo de Ciclo de Vida del Archivado Web". En *Anuario AC/E de cultura digital,* editado por AC/E Acción Cultural Española, 82-100, 2014. https://www.accioncultural.es/media/Default%20Files/activ/2014/Adj/Anuario_ACE_2014/7ArchivadoWeb_KHanna.pdf.

Harris, Verne. "The Archival Sliver: Power, Memory and Archives in South Africa". *Archival Science* 2, n.° 1-2 (2002): 63-86. DOI: 10.1007/BF02435631.

Hedstrom, Margaret. "Understanding Electronic Incunabula: A Framework for Research on Electronic Records". *The American Archivist* 54, n.° 3 (1 de julio de 1991): 334–354. https://doi.org/10.17723/aarc.54.3.125253r60389r011.

Hemphill, Libby, Margaret L. Hedstrom y Susan Hautaniemi Leonard. "Saving social media data: Understanding data management practices among social media researchers and their implications for archives". *Journal of the Association for Information Science and Technology* 72, n.º 1 (2021): 97–109. DOI: 10.1002/asi.24368.

Hernandez, Joe. "A judge sided with publishers in a lawsuit over the Internet Archive's online library". NPR. Publicado el 26 de marzo de 2023. https://www.npr.org/2023/03/26/1166101459/internet-archive-lawsuit-books-library-publishers.

Hockx-Yu, Helen. "Evaluating Twittervane". International Internet Preservation Consortium. https://netpreserve.org/projects/evaluating-twittervane/.

HTTrack WEBSITE COPIER. "Bienvenido". Consultado el 9 de junio de 2023. https://www.httrack.com/.

Ibero Ciudad de México. "A 10 años del origen del #YoSoy132 en la IBERO, su legado sigue vigente". Publicado el 9 de mayo de 2022. https://ibero.mx/prensa/10-anos-del-origen-del-yosoy132-en-la-ibero-su-legado-sigue-vigente.

Internet Archive. "About the Internet Archive". Consultado el 18 de mayo de 2023. https://archive.org/about/.

Internet Archive Global Events. "#RickyRenuncia web collection (Puerto Rico 2019)". Archive-It. Archivado desde julio de 2019. https://archive-it.org/collections/12491.

Internet Archive/Heritrix3. "Home". Consultado el 9 de junio de 2023. https://github.com/internetarchive/heritrix3/wiki.

Internet Archive. "Internet Archive Wayback Machine". https://archive.org/web/ y http://web.archive.org/.

_____. "Petabox". https://archive.org/web/petabox.php.

_____. "The stack: An introduction to the WARC file". https://ait.blog.archive.org/post/the-stack-warc-file/.

Internet Archive/Umbra. "Umbra". Consultado el 9 de junio de 2023. https://github.com/internetarchive/umbra.

International Internet Preservation Consortium. "About Archiving". Consultado el 5 de junio de 2023. https://netpreserve.org/web-archiving/about-archiving/.

_____. "Introducción". Última publicación el 6 de noviembre de 2007. https://archive-access.sourceforge.net/projects/wera/.

_____. "Novel Coronavirus (COVID-19)". Archive-It. Archivado desde febrero de 2020. https://archive-it.org/home/IIPC.

_____. "Tools and Software". IIPC. https://netpreserve.org/web-archiving/tools-and-software/.

IIPC/Twittervane. "Twittervane". https://github.com/iipc/twittervane.

Internet Live Stats. "Total number of websites". Consultado el 28 de junio de 2023. https://www.internetlivestats.com/total-number-of-websites/.

Instituto Nacional de Estadística y Geografía. "Estadísticas a propósito del Día Internacional de la Juventud". Comunicado de prensa 436 del INEGI, 10 de agosto de 2022. https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_Juventud22.pdf.

Invatati Afaceri. "Interfaz de programación de aplicaciones (API)". Publicado el 18 de noviembre de 2022. https://invatatiafaceri.ro/es/diccionario-financiero/interfaz-de-programacion-de-aplicaciones-api/.

Jackson, Gita. "The Geocities Archive is Bringing the Early Internet to Life". Vice. Publicado el 27 de enero de 2020. https://www.vice.com/en/article/n7jzgm/the-geocities-archive-is-bringing-the-early-internet-to-life.

Jackson, Nicholas y Alexis C. Madrigal. "The Rise and Fall of MySpace". *The Atlantic,* 12 de enero de 2011. https://www.theatlantic.com/technology/archive/2011/01/the-rise-and-fall-of-myspace/69444/.

Jules, Bergis, Ed Summers y Vernon Mitchell Jr. "Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities and Recommendations" (Documenting the Now White Paper), 1-12. Documenting the Now. Publicado en abril de 2018. https://www.docnow.io/docs/docnow-whitepaper-2018.pdf.

Kavvadia, Zefi. "An Overview of Social Media Archiving Tools" versión 1.0 (diciembre 2020), 15. Zenodo. Publicado el 2 de febrero de 2021. https://doi.org/10.5281/zenodo.4493594.

Kelleher, Christian. "Archives Without Archives: (Re) Locating and (Re) Defining the Archive Through Post-Custodial Praxis". *Journal of Critical Library and Information Studies* 1, n.° 2 (2017): 1-30. https://doi.org/10.24242/jclis.v1i2.29.

Kozlowski, Lori. "New Life: How MySpace Spawned a Start-Up Ecosystem". Forbes, 15 de mayo de 2012. https://www.forbes.com/sites/lorikozlowski/2012/05/15/how-myspace-spawned-a-startup-ecosystem/?sh=2dee0c6040ba.

Kreymer, Ilya. "A New Phase for Webrecorder Project, Conifer and ReplayWeb.page". Webrecorder Web archiving for all! Publicado el 11 de junio de 2020. https://webrecorder.net/2020/06/11/webrecorder-conifer-and-replayweb-page.html.

Kreymer, Ilya y Emma Dickson. "Announcing WACZ Format 1.0". Webrecorder Web archiving for all! Publicado el 18 de enero de 2021. https://webrecorder.net/2021/01/18/wacz-format-1-0.html.

Kussmann, Carol, NDSA, Katherine Kim, Bethany Nowviskie, Wayne Graham, Becca Quon, Winston Atkins, Aliya Reich, Matt Schultz, Lauren Work, Paige Walker, Nathan Tallmanet. "National Digital Stewardship Alliance (NDSA) / Web Archiving Survey". OSF. Publicado el 6 de enero de 2022. https://osf.io/4ytb2/.

Lampert, Cory y Emily Lapworth. "What do we mean by «collections as data» (CAD)?". UNLV University Libraries. Enviado el 2 de marzo de 2020. https://www.library.unlv.edu/whats-new-special-collections/2020/2020-03/what-do-we-mean-collections-data-cad-cory-lampert-emily#:~:text="Collections%20as%20data"%20is%20the,or%20people%20that%20are%20named.

Library of Congress. "ARC_IA, Internet Archive ARC file format". Última actualización 20 de abril de 2022. Consultado el 5 de junio de 2023. https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml.

_____. "Format Web Archive". https://www.loc.gov/web-archives/.

_____. "Library of Congress Recommended Formats Statement 2022-2023". Consultado el 16 de junio de 2023. https://www.loc.gov/preservation/resources/rfs/RFS%202022-2023-ArchivalOnly.pdf.

Library of Congress. "Library of Congress Recommended Formats Statement 2023-2024". Consultado el 16 de junio de 2023. https://www.loc.gov/preservation/resources/rfs/index.html.

_____. "Saving the World Wide Web". Consultado el 18 de mayo de 2023. https://www.digitalpreservation.gov/series/challenge/web_harvest_challenge.html.

_____. "Sustainability of Digital Formats: Planning for Library of Congress Collections". Consultado el 5 de junio de 2023. https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml.

_____. "Web Archive Collection Zipped". Consultado el 5 de junio de 2023. https://www.loc.gov/preservation/digital/formats/fdd/fdd000586.shtml.

_____. "X. Web Archive". Consultado el 16 de junio de 2023. https://www.loc.gov/preservation/resources/rfs/webarchives.html.

Literary Machines. "A wayback machine (pywb) on a cheap, shared host". Publicado el 24 de octubre de 2014. https://literarymachin.es/pywb-wayback-machine/.

Littman, Justin. "Web archiving and/or/vs social media API archiving". Social Feed Manager. Publicado el 13 de diciembre de 2017. https://gwu-libraries.github.io/sfm-ui/posts/2017-12-13-web-social-media-archiving.

Lyman, Peter y Hal Varian. "How Much Information? 2000". Universidad de California. Publicado el 18 de octubre de 2000. http://www.sims.berkeley.edu/research/projects/how-much-info/.

Lyman, Peter. "Problem Statement: Why Archive the Web?". Council on Library and Information Resources. Consultado el 18 de mayo de 2023. https://www.clir.org/pubs/reports/pub106/web/#1.

Masanès, Julien. "Web Archiving: Issues and Methods". En *Web Archiving,* editado por Julien Masanès, 1-45. Nueva York: Springer, 2006.

McMillan, Robert. "The Friendster Autopsy: How a Social Network Dies". *Wired,* 27 de febrero de 2013. https://www.wired.com/2013/02/friendster-autopsy/.

Milligan, Ian. "La historia en la era de la abundancia: archivos web e investigación histórica". *Historia y Memoria,* n.º especial, 10 años (2020): 235-269. https://doi.org/10.19053/20275137.nespecial.2020.11587.

Molina Suárez, María Jesús. "Archivo web de las publicaciones en línea en las comunidades autónomas". *Cartas diferentes. Revista canaria de patrimonio documental,* n.º 15 (2019): 283-307. https://mdc.ulpgc.es/s/mdc/item/125193.

Montilla Peña, Leomar José y Mayra M. Mena Mujica. "Estado de desarrollo de la archivística clásica hasta los años 30 del siglo XX: Tres manuales archivísticos de trascendencia universal". *Biblios Revista de Bibliotecología y Ciencias de la Información,* n.º 52 (octubre de 2013): 43-58. DOI: 10.5195/biblios.2013.122.

Mordell, Devon. "Critical Questions for Archives as (Big) Data". *Archivaria* 87 (mayo de 2019): 140-161. https://archivaria.ca/index.php/archivaria/article/view/13673.

Moreau, Elise. "Is MySpace Dead? The troubled social network's struggle to make a real comeback". Lifewire Tech for Humans. Actualizado el 21 de enero de 2022. https://www.lifewire.com/is-myspace-dead-3486012?utm_source=emailshare&utm_medium=social&utm_campaign=shareurlbuttons.

Moreno Freites, Zahira y Gertrudis Ziritt Trejo. "Redes Sociales como canales de digi-impacto en la participación ciudadana". *Utopía y Praxis Latinoamericana. Revista Internacional de Filosofía Iberoamericana y Teoría Social* 24, n.° 3 (noviembre 2019): 30-45. https://produccioncientificaluz.org/index.php/utopia/article/view/29683.

Mundt, Marcia, Ross Karen y Charla M. Burnett. "Scaling Social Movements Through Social Media: The Case of Black Lives Matter". *Social Media + Society* 4, n.° 4 (October-December 2018): 1-14. https://doi.org/10.1177/2056305118807911.

National Library of Korea OASIS. "OASIS (Online Archiving & Searching Internet Sources)". https://nl.go.kr/oasis/.

National Széchényi Library of Hungary. "SolrWayback". https://webadmin.oszk.hu/solrwayback/.

NetArchiveSuite/NetarchiveSuite. "About". Consultado el 9 de junio de 2023. https://github.com/netarchivesuite/netarchivesuite#readme.

Ngak, Chenda. "Then and now: a history of social networking sites". *CBS News,* 6 de julio de 2011. https://www.cbsnews.com/pictures/then-and-now-a-history-of-social-networking-sites/.

NWA. Consultado el 18 de mayo de 2023. https://nwatoolset.sourceforge.net/.

Odabas, Meltem. "10 facts about Americans and Twitter". Pew Research Center. Publicado el 5 de mayo de 2022. https://www.pewresearch.org/short-reads/2022/05/05/10-facts-about-americans-and-twitter/.

Oliver, Gillian y Ross Harvey. *Digital Curation*. 2.ª ed. Chicago: ALA Neal-Schuman, 2016.

Osterberg, Gayle. "Update on the Twitter Archive at the *Library of Congress". Library of Congress Blogs,* 26 de diciembre de 2017. https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/.

Oury, Clément. "WARC implementation guidelines. Contribution from WARC usage task force". IIPC. Publicado el 27 de enero de 2009. https://netpreserve.org/resources/WARC_Guidelines_v1.pdf.

Pandora Australia's Web Archive. "PANDORA Overview". Consultado el 19 de mayo de 2023. https://pandora.nla.gov.au/overview.html.

"PANDORA Digital Archiving System (PANDAS)". Pandora: Australia's Web Archive National Library of Australia and Partners. http://pandora.nla.gov.au/pandas.html.

Perma.cc. "About Perma.cc". Consultado el 9 de junio de 2023. https://perma.cc/.

Ponce, Isabel. "Monográfico: Redes Sociales". Observatorio Tecnológico. Publicado el 17 de abril de 2012. http://recursostic.educacion.es/observatorio/web/en/internet/web-20/1043-redes-sociales.

Quiroga, Nicolás. "Interpretación histórica y objetos digitales: consideraciones a partir de ejemplos concretos". *Vegueta. Anuario de la Facultad de Geografía e Historia* 22, n.º 1 (2022): 39-55. https://doi.org/10.51349/veg.2022.1.03.

R3D Red en Defensa de los Derechos Digitales. "Editoriales exigen cierre del Internet Archive en demanda por préstamo de libros digitalizados". R3D. Publicado el 23 de marzo de 2023. https://r3d.mx/2023/03/23/editoriales-exigen-cierre-del-internet-archive-en-demanda-por-prestamo-de-libros-digitalizados/.

Rauber, Andreas, Max Kaiser y Bernhard Wachter. "Ethical issues in Web Archive Creation and Usage – Towards a Research Agenda". En *8th International Web Archiving Workshop (IWAW08),* 2008. https://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda.

Reynolds, Emily. *Web Archiving Uses Cases.* Library of Congress, UMSI, ASB13, March 7, 2013, 1-10. https://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf.

Rising, David y Barbara Ortutay. "Germany to Facebook: Stop forcing users to share their data". *The Associated Press,* 7 de febrero de 2019. https://apnews.com/article/ap-top-news-facebook-privacy-scandal--social-platforms-germany-north-america-04440c1ca08b4caf9da2f6e9bf0038d7.

Rodríguez Reséndiz, Perla Olivia. "La preservación digital sonora". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 30, n.° 68 (2016): 173-195. https://doi.org/10.1016/j.ibbai.2016.02.009.

Rodríguez Reséndiz, Perla Olivia. *Estado de la preservación digital en los archivos sonoros y audiovisuales de Iberoamérica*. Quito: Universidad Andina Simón Bolívar, Sede Ecuador, 2020. https://www.cyted.org/sites/default/files/estado_de_preservacion_19enero2021_1_0.pdf.

_____, Joséphine Simonnot y Dafne Citalli Abad Martínez. "Gestor de contenidos de código abierto para archivos digitales sonoros que preservan materiales de investigación". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32, n.° 77 (2018): 101-115. http://dx.doi.org/10.22201/iibi.24488321xe.2018.77.58005.

Rosenberg, Matthew, Nicholas Confessore y Carole Cadwalladr. "How Trump Consultants Exploited the Facebook Data of Millions". *The New York Times,* 17 de marzo de 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

Ros-Martín, Marcos. "Evolución de los servicios de redes sociales en internet". *El profesional de la información* 18, n.° 5 (septiembre-octubre de 2009): 552-557. DOI: 10.3145/epi.2009.sep.10.

Ruiz Gómez, Vicenç y Aniol Maria Vallès... "#Cuéntalo: the path between archival activism and the social archive(s)". *Archives & Manuscripts. The Journal of the Australian Society of Archivists* 48, n.º 3 (2020): 271–290. https://doi.org/10.1080/01576895.2020.1802306.

Sancho, Guiomar Rovira. "El #YoSoy 132 mexicano: la aparición (inesperada) de una red activista / The Mexican #YoSoy 132: the (unexpected) emergence of a activist network". *Revista CIDOB d'Afers Internacionals,* n.º 105 (abril de 2014): 47-66. https://www.cidob.org/es/articulos/revista_cidob_d_afers_internacionals/105/el_yosoy132_mexicano_la_aparicion_inesperada_de_una_red_activista.

Saving Ukrainian Cultural Heritage Online. "About SUCHO", SUCHO, https://www.sucho.org/about.

Sheffield, Rebecka Taves. "Facebook Live as a Recordmaking Technology". *Archivaria* 85 (mayo 2018): 96-121. https://archivaria.ca/index.php/archivaria/article/view/13632.

Singer, Natasha. "What You Don't Know About How Facebook Uses Your Data". *The New York Times,* 11 de abril de 2018. https://www.nytimes.com/2018/04/11/technology/facebook-privacy-hearings.html.

"Social.coop". Mastodon. https://social.coop/@edsu.

Social Feed Manager. "Social Feed Manager Helping researchers and archivists build social media collections". Consultado el 9 de junio de 2023. https://gwu-libraries.github.io/sfm-ui/.

"State of the WARC Report: Web archive management and preservation in 2019-20". *Archive-It*, 26 de mayo de 2020. https://ait.blog.archive.org/post/state-of-the-warc-2020/ y https://archive-it.org/blog/files/2020/05/State-of-the-WARC_2020.pdf.

Statista Research Department. "Number of Twitter users worldwide from 2019 to 2024". Statista. Publicado el 14 de diciembre de 2022. https://www.statista.com/statistics/303681/twitter-users-worldwide/.

_____. "Las redes sociales en México – Datos estadísticos". Statista. Publicado el 27 de marzo de 2023. https://es.statista.com/temas/7392/las-redes-sociales-en-mexico/#topicOverview.

Statista Research Department. "Número de usuarios de redes sociales en México de 2017 a 2027". Statista. Publicado el 27 de marzo de 2023. https://web.archive.org/web/20230523033503/https://es.statista.com/estadisticas/1141228/numero-de-usuarios-de-redes-sociales-mexico/.

_____. "Porcentaje de usuarios por red social en México en 2022". Statista. Publicado el 28 de marzo de 2023. https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/.

Stirling, Peter y Jaanus Kõuts. "How to fit in? Integrating a web archiving program in your organization". Workshop report and evaluation. Framework Education and Training BnF, Paris, France, November 26-30 2012. IIPC. Publicado en abril de 2013. https://netpreserve.org/resources/IIPC_project-Report_and_Evaluation_of_BnF_IIPC_Workshop_on_Web_Archiving.pdf.

Summers, Ed. "A Ferguson Twitter Archive". Inkdroid. Publicado el 30 de agosto de 2014. https://inkdroid.org/2014/08/30/a-ferguson-twitter-archive/.

_____. "Appraisal Talk in Web Archives". *Archivaria: The Journal of the Association of Canadian Archivists* 89, n.º 1 (mayo de 2020): 70-103. https://archivaria.ca/index.php/archivaria/article/view/13733.

_____. "Looking Forwards". Medium. Publicado el 19 de julio de 2023. https://news.docnow.io/looking-forwards-64cee8436640.

Summers, Ed y Ricardo Punzalan. "Bots, Seeds and People: Web Archives as Infrastructure". En *CSCW'17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing,* 821-834. New York, USA: Association for Computing Machinery, 25 de febrero de 2017. https://doi.org/10.1145/2998181.2998345.

The Guardian. "The Cambridge Analytica Files". Consultado el 7 de julio de 2023, https://www.theguardian.com/news/series/cambridge-analytica-files.

Thibodeau, Kenneth. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years". En *The State of Digital Preservation: An International Perspective,* 4-31. Washington, D. C.: Council on Library and Information Resources, 2002. http://www.clir.org/pubs/abstract/pub107abst.html.

Thomson, Sara Day. *Preserving Social Media*. Great Britain: Digital Preservation Coalition, 2016. http://dx.doi.org/10.7207/twr16-01.

Thylstrup, Nanna Bonde. "La memoria digital del mundo está en peligro". *The New York Times,* 25 de junio de 2023, párr. 6. https://www.nytimes.com/es/2023/06/25/espanol/opinion/internet-archivo-digital.html?smid=nytcore-ios-share&referringSource=articleShare.

Tpadilla. "Build, Access, Analyze: Introducing ARCH (Archives Research Compute Hub)". *Internet Archive Blogs,* 26 de junio de 2023. https://blog.archive.org/2023/06/26/build-access-analyze-introducing-arch-archives-research-compute-hub/.

Treem, Jeffrey W., Stephanie L. Dailey, Casey S. Pierce y Diana Biffl. "What We Are Talking About When We Talk About Social Media: A Framework for Study". *Sociology Compass* 10, n.º 9 (2016): 768–784. DOI: 10.1111/soc4.12404.

Trove. "Website category. Restricted content". https://trove.nla.gov.au/help/categories/websites-category.

Trove. "Websites". Consultado el 19 de mayo de 2023. https://trove.nla.gov.au/search/category/websites?keyword=1991.

Truman, Gail. *Web Archiving Environmental Scan.* Harvard Library Report, 2016. https://dash.harvard.edu/handle/1/25658314.

Tsutomu, Shimura. "20 Years of the Web Archiving Project (WARP) at the National Diet Library, Japan". International Internet Preservation Consortium. Publicado el 3 de abril de 2023. https://netpreserveblog.wordpress.com/2023/04/03/20-years-of-the-web-archiving-project-warp-at-the-national-diet-library-japan/.

"Twitter Archives". University of Michigan Library. https://deepblue.lib.umich.edu/handle/2027.42/116594.

UKWA UK Web Archives. "Frequently asked questions". https://www.webarchive.org.uk/en/ukwa/info/faq/#how-frequently-are-websites-collected y https://www.webarchive.org.uk/en/ukwa/info/faq/#how-big-is-the-archive.

_____. "Topics and Themes". https://www.webarchive.org.uk/en/ukwa/category/.

Unesco. *Convirtiendo la amenaza del COVID-19 en una oportunidad para un mayor apoyo al patrimonio documental.* Unesco. Publicado el 5 de abril de 2020. https://en.unesco.org/sites/default/files/dhe-covid-19-unesco_statement_es.pdf y https://www.unesco.org/en/articles/turning-threat-covid-19-opportunity-greater-support-documentary-heritage.

_____. *Directrices Unesco/PERSIST sobre selección del patrimonio digital para su conservación a largo plazo.* 2.ª ed., mayo de 2021. https://repository.ifla.org/bitstream/123456789/1390/1/UNESCO%20PERSIST%20sobre%20selecci%C3%B3n%20del%20patrimonio%20digital%20para%20su%20conservaci%C3%B3n%20a%20largo%20plazo%20-%202nd%20edici%C3%B3n.pdf.

_____. "Convención para la Salvaguardia del Patrimonio Cultural Inmaterial". Reunión 32 Conferencia General de la Organización de las Nacionales Unidas para la Educación, la Ciencia y la Cultura, celebrada en París, del 29-septiembre al 16- octubre de 2003. Unesco Patrimonio Cultural Inmaterial. https://ich.unesco.org/es/convenci%C3%B3n.

_____. *Memoria del Mundo. Directrices para la salvaguardia del patrimonio documental,* edición revisada y preparada por Ray Edmondson. París: Unesco, 2002). https://unesdoc.unesco.org/ark:/48223/pf0000125637_spa.

_____ *Recomendación sobre la Salvaguardia y la Conservación de las Imágenes en Movimiento.* Reunión 21 Conferencia General de la Organización de las Nacionales Unidas para la Educación, la Ciencia y la Cultura, celebrada en Belgrado, del 23-septiembre al 28-octubre de 1980. Unesco. https://www.unesco.org/es/legal-affairs/recommendation-safeguarding-and-preservation-moving-images.

Van Dijck, José. *La cultura de la conectividad: una historia crítica de las redes sociales.* Buenos Aires: Siglo xxi Editores, 2016.

Vosoughi, Soroush, Deb Roy y Sinan Aral. "The Spread of True and False News Online". mit Initiative on the Digital Economy. Publicado en marzo de 2018, (sección Publications/Research Briefs), 1-5. https://ide.mit.edu/wp-content/uploads/2018/12/2017-IDE-Research-Brief-False-News.pdf.

VR, Ferose. "Digital Librarian for and of the World". Medium. Publicado el 14 de junio de 2023. https://ferosevr.medium.com/digital-librarian-for-of-the-world-9ec7cf1a239e.

W3 Schools. "json – Introduction". Consultado el 25 de enero de 2023. https://www.w3schools.com/js/js_json_intro.asp.

Web Curator Tool. "Overview and History". https://webcuratortool.readthedocs.io/en/latest/guides/overview-history.html#introduction.

Wikipedia. "List of Web archiving initiatives". Última modificación el 10 de junio de 2023. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

_____. "Internet Archive". Última modificación el 27 de octubre de 2023. https://en.wikipedia.org/wiki/Internet_Archive#cite_note-86.

Williams, Nate. "The Real Reason MySpace Failed Spectacularly". History-Computer. Publicado el 13 de diciembre de 2022. https://history-computer.com/the-real-reason-my-space-failed-spectacularly/#:~:text=At%20its%20prime%2C%20MySpace%20was,-mass%20shift%20to%20new%20platforms.

Yakel, Elizabeth. "Digital curation". OCLC *Systems & Services: International digital library perspectives* 23, n.º 4 (2007): 335-340. https://doi.org/10.1108/10650750710831466.

Young, Tyler A. "General overview". IIPC/OpenWayback. Editado el 24 de julio de 2018. https://github.com/iipc/openwayback/wiki/General-overview.

Zimmer, Michael. "The Twitter Archive at the Library of Congress: Challenges for information practice and information policy". *First Monday* 20, n.º 7 (6 de julio de 2015). https://doi.org/10.5210/fm.v20i7.5619.

Zuboff, Shoshana. *The Age of Surveillance Capitalism.* New York: Public Affairs, 2018.

Preservación digital de contenidos publicados en la web y las redes sociales. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Sergio Sepúlveda; revisión especializada, Angélica Valenzuela; revisión de pruebas, Carlos Cevallos Sosa y Sergio Sepúlveda; formación editorial, Oscar Fernando Arcos Casañas y Mónica Salmorán Alvarado. Fue impreso en papel cultural de 90 g en los talleres de Kronos Digital S. A. de C. V., 5 de febrero 436-B, Colonia Algarin, C. P. 06880, Alcaldía Cuauhtémoc, Ciudad de México. Se terminó de imprimir en febrero 2024.