Historia y MEMORIA
ISSN: 2027-5137
Universidad Pedagógica y Tecnológica de Colombia (UPTC)

Milligan, Ian
La historia en la era de la abundancia: archivos web e investigación histórica*
Historia y MEMORIA, Esp., 2020, Enero-Diciembre, pp. 235-269
Universidad Pedagógica y Tecnológica de Colombia (UPTC)

DOI: https://doi.org/10.19053/20275137.nespecial.2020.11587

Disponible en: https://www.redalyc.org/articulo.oa?id=325166074007

ED
ESPECIAL

Historia Y MEMORIA

ISSN: 2027-5137 Número Especial • 10 Años • Año 2020 - Tunja, Colombia

La historia en la era de la abundancia: archivos web e investigación histórica

https://doi.org/10.19053/20275137.nespecial.2020.11587

Ian Milligan

Páginas 235-269

La historia en la era de la abundancia: archivos web e investigación histórica*

Ian Milligan[1]
University of Waterloo - Canadá

Recepción: 05/02/2020
Evaluación: 11/06/2020
Aprobación: 1/07/2020

Artículo de Investigación e Innovación

https:/doi.org/10.19053/20275137.nespecial.2020.11587

Resumen

¿La década de 1990 puede ser considerada historia del presente? A medida que los historiadores se dediquen a estudiar este período y los sucesivos, se encontrarán con un registro histórico que es radicalmente diferente de lo que ha existido anteriormente. Las webs antiguas, las redes sociales, los blogs, las fotografías y los videos son parte de la abrumadora cantidad de información digital que los tecnólogos, bibliotecarios, archivistas y organizaciones como Internet Archive han estado recopilando durante las últimas tres décadas. El artículo explora cómo esta significativa variación en nuestro registro histórico cambió el trabajo de los historiadores. Lo hace de dos maneras fundamentales. Primero, describe los enfoques, métodos, herramientas y funciones de búsqueda que pueden ayudar a un historiador a convertir documentos web en fuentes históricas. En segundo lugar, considera las implicaciones del tamaño y la escala de las fuentes digitales, que equivalen a más información de la

* Traducción del artículo realizada por el Dr. Anaclet Pons.

1 Doctor en Historia por la Universidad de York profesor asociado de historia en la Universidad de Waterloo. Últimas publicaciones: *History in the Age of Abundance? How the Web is Transforming Historical Research* (Montreal & Kingston: McGill-Queen's University Press, 2019); Niels Brügger e Ian Milligan, eds., *SAGE Handbook of Web History* (London: SAGE, 2018). ✉ i2milligan@uwaterloo.ca https://orcid.org/0000-0002-1470-7723.

que los historiadores jamás han tenido al alcance de la mano, mucha de la cual es de personas que tradicionalmente han estado ausentes del registro histórico. Estos aspectos se han hecho tangibles a través de un minucioso estudio de caso, trabajando con el archivo web de GeoCities.com, una colección de cientos de millones de páginas web de la década de 1990.

Palabras clave: historia digital, archivos web, historiografía, humanidades digitales.

History in the era of abundance: web archives and historical research

Abstract

Can the 1990s be considered the history of the present? As historians turn to study this period and beyond, they will encounter a historical record that is radically different from what has ever existed before. Old websites, social media, blogs, photographs, and videos are all part of the massive quantities of digital information that technologists, librarians, archivists, and organizations such as the Internet Archive have been collecting for the past three decades.

This article explores how this dramatic shift in our historical record changes the work of historians. It does so in two main ways. First, it outlines the approaches, methods, tools, and search functions that can help a historian turn web documents into historical sources. Secondly, it considers the implications of the size and scale of digital sources, which amount to more information than historians have ever had at their fingertips, and much of which are by and about people who have traditionally been absent from the historical record. As a way to make these points tangible, it does so primarily through an in-depth case study of working with the GeoCities.com web archive, a collection of hundreds of millions of 1990s webpages.

Key Words: digital history, web archives, historiography, digital humanities.

L'histoire à l'ère de l'abondance: archives web et recherche historique

Résumé

Les années peuvent être considérées comme faisant partie de l'histoire du présent? Au fur et à mesure que les historiens étudient cette période, ils feront face à des archives radicalement autres par rapport à celles du passé. La web, les réseaux sociaux, les blogs, les photos et les vidéos font désormais partie de la quantité d'information numérique que de nombreux professionnels et des organisations comme Internet Archive se chargent de rassembler il y a déjà 30 ans. Cet article explore la manière dont cette importante variation de nos registres historiques a changé le travail des historiens, et ce en deux sens: premièrement, on décrit les approches, les méthodes, les outils et les fonctions de recherche qui peuvent aider les historiens à convertir les documents web en sources historiques; deuxièmement, on considère les implications de la taille et de l'échelle des sources numériques, bien plus abondante que celle dont les historiens ont disposé auparavant, des sources concernant des individus qui sont traditionnellement absents du registre historique. Ces aspects-là sont devenus plus concrets grâce à une étude de cas: GeoCities.com, une collection de plusieurs centaines de millions de sites internet des années 1990.

Mots-clés: histoire numérique, archives web, historiographie, humanités digitales.

Introducción

Desde el advenimiento de la World Wide Web en 1990, el registro histórico ha experimentado un cambio impresionante. Los archivos tradicionales estaban y están limitados por severas restricciones de espacio y por el largo proceso a partir del cual las evidencias históricas se adquieren, documentan y ponen a disposición de los investigadores. Hoy en día, cualquier persona con conexión a internet puede seleccionar y preservar casi instantáneamente un sitio web o un tuit. Y esto ha estado

sucediendo a una escala impresionante. Desde 1996, Internet Archive ha reunido unos cuarenta petabytes –un petabyte son mil terabytes, que a su vez son mil gigabytes– de datos, y otras bibliotecas nacionales de todo el mundo probablemente han recopilado también lo mismo. En resumen, el actual registro cultural digital –unos ochenta petabytes de contenido– forma un cuerpo de información histórica como nunca antes habíamos visto.

Este nuevo registro histórico representa un cambio contundente en dos aspectos clave. En primer lugar, la *escala* de los datos: los cuarenta petabytes de Internet Archive solo comprenden alrededor de 635 mil millones de capturas únicas de webs. Estos documentos son parte de otros cientos de miles de millones (y pronto billones, ya que esto se duplica cada dos o tres años), algo que representa un incremento notable con respecto a lo que estamos acostumbrados. En segundo lugar, el alcance de estos datos se ha expandido: ahora se están compendiando datos que nunca se hubieran recopilado, acerca de personas que no estaban tradicionalmente en el registro histórico: páginas web para niños, páginas web de movimientos sociales, videos aleatorios de YouTube, etcétera.

Este artículo explora el cambio en el medio sobre el que se sustentan los archivos web, argumentando que nuestro registro histórico está siendo y quedará profundamente afectado por los archivos web y otros tipos similares de repositorios digitales. Esta es un arma de doble filo: ampliar el alcance y la escala abre nuevas oportunidades, pero estos mismos factores acarrean desafíos considerables. Trabajar con terabytes de datos históricos no es sencillo. Aprovechando la publicación reciente del libro de mi autoría, *History in the Age of Abundance?*[2]. Este artículo se propone tres objetivos principales: primero, presentar los conceptos básicos que los historiadores deben comprender para usar los archivos web; segundo, explorar las implicaciones del cambio en alcance y escala; y, en tercer lugar, proporcionar algunas vías concretas

2 Ian Milligan, *History in the Age of Abundance? How the Web is Transforming Historical Research* (Montreal & Kingston: McGill-Queen's University Press, 2019).

sobre cómo los historiadores pueden usar los archivos web. En otras palabras, cómo pueden los historiadores afrontar este desafío.

1. No es un asunto secundario: archivos web e investigación histórica

Los archivos web no son una preocupación secundaria ni algo para determinado grupo de interesados. Algunos ejemplos pueden ilustrarlo. Los medios tradicionales llegan cada vez más a la mayor parte de su público a través de webs, las cuales en algunos casos pueden evolucionar en línea y diferir drásticamente de sus contrapartes impresas. Un historiador que estudie el siglo XXI, por ejemplo, no puede confiar en la edición impresa de un periódico; ello implicaría omitir cómo la mayoría de las personas han hecho uso de los medios de comunicación en la última década. Tanto las pequeñas como las grandes empresas utilizan webs para comercializar y vender sus productos, así como para atraer inversores y concitar adhesión hacia las marcas corporativas. Los departamentos gubernamentales, en todos los niveles, involucran al público a través de sus sitios web, proporcionando información sobre servicios, problemas y los objetivos de sus políticas. El seguimiento de cómo han cambiado con el tiempo puede arrojar luz sobre la naturaleza de la administración pública y sobre cómo implementan las políticas. Los partidos políticos utilizan sitios y plataformas web para llegar al público en general, durante las campañas electorales en diferentes momentos. Los activistas y las organizaciones de justicia social pueden reunir a amplias comunidades en línea, conformando un tesoro oculto de información para historiadores y académicos que busquen comprender los movimientos sociales. De hecho, escribo esto durante el pedido de «quédese en casa» por el COVID-19 -casi toda mi existencia, y la de muchos de mis colegas, la estoy viviendo mediada por la tecnología, lo que produce un archivo exhaustivo.

Gran parte de esto sigue siendo especulativo. El número de historiadores que llevan a cabo investigaciones posteriores a 1996 es aún bastante pequeño; sin embargo,

no hace falta mucha imaginación para darse cuenta de que pronto estos archivos web transformarán la forma en que llevamos a cabo nuestros estudios. Ningún historiador puede ignorar la fenomenal transformación de la información y de los registros de experiencia en línea, lo que subraya la necesidad de que la profesión histórica se implique con el impacto de los archivos web. Descuidar la web supondría ignorar el medio principal de comunicación, de publicación, de interacción social, de comercio empresarial y de actividad creativa desde la década de 1990. Eso no quiere decir que este sea un registro mágicamente completo de la sociedad. El archivo web no producirá un registro «completo» de nuestro mundo, no vivimos en un Estado de vigilancia que todo lo ve. Ningún registro archivístico puede capturar completamente la humanidad, la riqueza y la complejidad esenciales de nuestra existencia, de absolutamente todo lo que nos envuelve. Pero, aun así, los archivos web representan un cambio radical, ya que nuestro registro histórico –si bien aún incompleto– es exponencialmente más grande que nunca.

2. La historia en la era del algoritmo

Dado que ningún ser humano puede leer ni siquiera una pequeña fracción de las páginas de un archivo web de tamaño medio, tendremos que confiar en las computadoras para dar sentido a los datos. En el pasado, los historiadores confiaban en los archivistas y en los bibliotecarios para administrar y organizar los documentos de archivo; sin embargo, la magnitud actual de estos datos indica que los historiadores tendremos que realizar este arduo trabajo por nuestra cuenta. Esto quiere decir que los historiadores necesitaremos desarrollar habilidades técnicas para organizar datos y encontrar documentos relevantes e importantes dentro del archivo. Esto requiere habilidades muy diferentes de las asociadas a la consulta de los repositorios tradicionales en papel. Desafortunadamente, hoy no resulta nada fácil que las habilidades necesarias para usar las fuentes digitales se obtengan en un programa de grado o de posgrado en historia.

Ahora bien, si los historiadores quieren acceder a esta información en la era del algoritmo, entonces la alfabetización digital se está convirtiendo en algo crucial. Los algoritmos se encuentran en el centro de casi todas las investigaciones con recursos digitales, desde los archivos web hasta la prensa histórica digitalizada[3]. Esto es algo que vemos a diario cuando ejecutamos nuestras propias búsquedas en Google y rara vez avanzamos más allá de la primera página de resultados. Es posible que Google haya indexado diversas voces, pero si su algoritmo las asigna a la página quinta, la quincuagésima o a la que hace quinientos de sus resultados, las voces también podrían ser inexistentes[4]. La capacidad de descubrir algo es importante. Cada vez que usamos un motor de búsqueda se toma una decisión entre bambalinas para mostrar un resultado determinado como # 1 –y otro como # 1.000.000–, asegurando virtualmente que no hagamos clic en este último. Como mínimo, los historiadores necesitamos estar al corriente sobre el modo en que estos algoritmos dan forma a nuestro trabajo.

A medida que los humanistas dominan y exploran las implicaciones del giro digital, se hace evidente que las preguntas sobre «cómo analizar, interpretar y explotar los datos masivos son grandes problemas para las humanidades»[5]. A la luz del inminente diluvio de archivos web de origen digital, son especialmente los historiadores quienes necesitan convertirse en líderes dentro de las humanidades digitales. Los objetos de investigación desde la década de 1990 en adelante lo requieren, y pronto –si no ha ocurrido ya– se asentarán firmemente en el terreno de la historia. Los historiadores están entrando en la era del algoritmo. ¿Pero cómo será esto? En la siguiente sección, me propongo mostrar cómo estas preocupaciones abstractas pueden hacerse tangibles, a través de un extenso

3 Lara Putnam, «The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast,» *The American Historical Review,* n° 121.2 (2016): 377–402; Shawn Graham, Ian Milligan y Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (Lonres: Imperial College Press, 2015), 2015.

4 Siva Vaidhyanathan, *The Googlization of Everything (And Why We Should Worry)* (Berkeley: University of California Press, 2011).

5 Geoffrey Rockwell y Stéfan Sinclair, *Hermeneutica: Computer-Assisted Interpretation in the Humanities* (Cambridge, MA: The MIT Press, 2016), 14.

estudio de caso de investigación histórica, centrado en un archivo web de los años noventa.

3. Las implicaciones del alcance y la escala: el caso de GeoCities

Uno de los sitios web que he usado en muchas de mis investigaciones anteriores es GeoCities.com[6]. Fundada a fines de 1994, GeoCities permitió a los usuarios web crear rápida y fácilmente sus propias webs de forma gratuita. A medida que la web se hizo popular a mediados de la década de 1990, el sitio experimentó un aumento meteórico en el número de usuarios. Las personas no solo querían navegar por la web para ver el contenido de otras personas, sino también para agregar sus voces a la mezcla. A mediados de 1998, conforme se propagaba el entusiasmo por la web, cada día 18.000 nuevos usuarios creaban cuentas y webs[7]. En 1999, llegó a ser el tercer sitio más visitado de toda la web, abarcando finalmente al menos 186 millones de «documentos», repartidos en más de siete millones de páginas web[8]. Hoy, sin embargo, si alguien visita GeoCities.com solo encontrará un anuncio del servicio de alojamiento de webs de Yahoo! Ningún rastro de la comunidad en tiempos vibrantes permanece en ese sitio, ni lo ha hecho desde su cierre en 2009.

El alcance y la escala de GeoCities significan que puede servirnos como un microcosmos de los desafíos que se han mencionado anteriormente respecto a los archivos web.

6 Ian Milligan, «Welcome to the Web: The Online Community of GeoCities and the Early Years of the World Wide Web,» en *The Web as History*, ed. Niels Brügger y Ralph Schroeder (Londres: UCL Press, 2017); Ian Milligan, «Exploring Web Archives in the Age of Abundance: The Case of GeoCities,» en *SAGE Handbook of Web History*, ed. Niels Brügger y Ian Milligan (Londres: SAGE, 2018).

7 John Motavalli, *Bamboozled at the Revolution: How Big Media Lost Billions in the Battle for the Internet* (Nueva York; Penguin Group, 2004), 191.

8 Es difícil saber con certeza cuán grande era GeoCities. La propia compañía tenía motivos para inflar números, y actualmente los archivos web no están completos. Sin embargo, podemos hallar citada en diversos lugares la cifra de siete millones, como en Dan Fletcher, «Internet Atrocity! GeoCities' Demise Erases Web History», *Time*, Nueva York, 9 de noviembre de 2009, http://content.time.com/time/business/article/0,8599,1936645,00.html. El número de «documentos» proviene de la cantidad de documentos HTML encontrados en un análisis con la plataforma warcbase de lo rastreado por Internet Archive con respecto al año 2009.

Lo que tenemos dentro de GeoCities es la evidencia de una comunidad en línea popular que nació, prosperó y declinó entre 1994 y 2009. Así pues, podemos decir que, por ser un lugar lleno de webs de personas comunes, GeoCities es un estudio de caso donde podemos ver si los archivos web podrían realizar potencialmente el sueño de una historia social más democrática.

Esto se debe a que, a medida que la web se introdujo en la cultura dominante de la sociedad norteamericana, entre mediados y finales de la década de 1990, GeoCities fue uno de los primeros sitios en acoger a los usuarios y facilitarles sus primeros pasos en la web. Por primera vez, los usuarios podían crear sus propias páginas web sin necesidad de tener habilidades de codificación, ni siquiera especiales destrezas técnicas. No es una hipérbole observar que GeoCities ayudó a que muchos nuevos usuarios pudieran acceder a la creación de contenido en la web. No quedarían meramente relegados a «navegar» por la web, sino que también podrían participar agregando a ella. De hecho, a eso podemos atribuir el descomunal crecimiento de GeoCities. Cinco semanas después de la apertura de GeoCities, se habían producido más de 600.000 «visitas» y, a finales de 1995, ya se habían creado 1.400 webs[9]. El primer hito, el de los 10.000 usuarios, se alcanzó en octubre de 1995, a los primeros 100.000 se llegó en agosto de 1996 y el primer millón de usuarios se logró en octubre de 1997. A mediados de 1998, el sitio era sin duda uno de los diez primeros en la web y crecía al ritmo de 18.000 nuevos usuarios al día[10]. Luego cayó casi tan rápido como había subido. En 1999, Yahoo! compró GeoCities por 4.6 mil millones de dólares. La cultura corporativa de la organización se transformó y, por razones que están fuera del alcance de este artículo, la prominencia de GeoCities disminuyó rápidamente, lo que llevó a su cierre en 2009.

9 Business Wire, «Beverly Hills Internet, Builder of Interactive Cyber Cities, Launches 4 More Virtual Communities Linked to Real Places,» *Business Wire*, 5 de julio de 1995, https://web.archive.org/web/20081211170054/http://www.allbusiness.com/marketing-advertising/marketing-advertising/7191644-1.html.

10 Motavalli, *Bamboozled at the Revolution*, 191.

4. Comunidad de lectura distante: la pregunta de investigación del estudio de caso

Una pregunta histórica clave sobre GeoCities es: ¿los usuarios encontraron allí una comunidad? Esto afecta a un debate más amplio sobre la naturaleza de las relaciones y de las conexiones personales en la era digital; volveré sobre algunas de estas cuestiones en breve.

La parte «geo» de GeoCities proviene de su enfoque espacial virtual: los sitios se agruparon en «vecindarios» o secciones del sitio basadas en temas. Los que escribían sobre «educación, literatura, poesía, filosofía» serían alentados a establecerse en Athens; los entusiastas de la política en CapitolHill; los pequeños empresarios o quienes trabajaban desde su casa en Eureka; etcétera. Algunos vecindarios incluían ciertas restricciones y unas guías explícitas, como las protectoras de EnchantedForest para niños. Otros tenían un alcance más amplio, como el vecindario más abultado, «Heartland», que se centraba en «familias, mascotas, valores de la localidad natal». Cada vecindario se presentaba como un mapa: el Enchanted Forest tendría una serie de cabañas, por ejemplo, a las que un usuario podría «mudarse» para alojar su sitio. A cada localidad se le asignaba un número de cuatro dígitos. Dentro de esa dirección de cuatro dígitos, los usuarios podían crear tantas páginas como pudieran caber dentro de su limitada capacidad de uno o dos megabytes (dependiendo de cuándo se registraron). A fines de 1996, había veintinueve vecindarios.

Cada vecindario reclutaba activamente nuevos «hacendados» (*homesteaders*) y ofrecía, a modo de ejemplo, tanto listas de temas sobre los que los usuarios podían escribir como otros sitios de vecindarios a los que podían acudir para leer y tener una idea del área. Los vecindarios populares comenzaron a llenarse rápidamente, lo que exigió la expansión hacia los «suburbios». Los usuarios pronto tuvieron que mudarse a barrios como Heartland/Plains o Heartland/Hills. Cada vecindario estaba limitado a 9000 sitios (1000-9999); Heartland, el vecindario más grande, llegó a tener 41

suburbios en 1999. Si bien la ampliación presentaba desafíos, GeoCities estaba comprometido con el enfoque espacial y temático. A este respecto, su objetivo declarado era fomentar comunidad.

La pregunta central es: ¿GeoCities tuvo éxito en fomentar comunidad? La comunidad era un aspecto central tanto de la retórica del marketing corporativo como de los objetivos establecidos, pero ¿lo experimentaron los usuarios? Como historiador, la comunidad puede ser difícil de definir y encontrar. Howard Rheingold ha impulsado una definición de las comunidades virtuales como «agregaciones sociales que emergen de la red cuando un número suficiente de personas entablan discusiones públicas durante un tiempo lo suficientemente largo, con suficiente sentimiento humano, para formar redes de relaciones personales en el ciberespacio» Él describió en particular el surgimiento de una economía basada en regalos, donde las personas ceden su tiempo sin recompensa directa (aunque, tal vez, en el futuro alguien les ayude)[11]. No basta con declarar simplemente que la comunidad existe, en una página de bienvenida de la web o en un comunicado de prensa, debe ser promulgada, recibida y percibida como tal por los miembros. En otras palabras, debemos encontrar evidencia de comunidad, no solo la palabra. Esto marca un uso ideal de los archivos web, el hecho de observar el comportamiento del usuario.

Se ha debatido si GeoCities era una comunidad virtual o no. La facilidad para unirse a GeoCities ha llevado a algunos académicos a descartar la noción de que fuera de entrada una comunidad. Christos J. P. Moschovitis ha argumentado que ofrecer simplemente alojamiento web y correo electrónico no era suficiente, señalando que muchos «miembros se inscribieron para tener espacio web gratuito, no para hacer nuevos amigos»[12]. Ciertamente, muchos usuarios de GeoCities

11 Howard Rheingold, *The Virtual Community: Homesteading on the Electronic Frontier* (Cambridge, Mass.: MIT Press, 2000), edición en digital, http://www.rheingold.com/vc/book/intro.html.

12 Christos J. P. Moschovitis, *History of the Internet: A Chronology, 1843 to the Present* (Santa Barbara, Calif.: ABC-CLIO, 1999).

hicieron exactamente eso: se registraron y crearon webs sin interactuar con sus vecinos digitales; sin embargo, hay rastros de comunidad virtual en el archivo web. Los nuevos usuarios fueron generosamente recibidos por los usuarios existentes, que tenían como objetivo facilitar sus primeros pasos en la web: enviaban un carrito de bienvenida para ayudarlos a aprender los entresijos del HTML, ofrecían consejos sobre el mantenimiento del sitio y, en algunos casos, «pasaban por allí» para dejar un comentario en el libro de visitas. Muchos otros se aseguraban de que sus conocimientos sobre los recursos del HTML estuvieran disponibles para su consulta. Con ese talante, se alentaba a los usuarios a comunicarse con sus nuevos vecinos, y se esperaba que hicieran «apariciones» regulares mejorando continuamente su web. Nada de esto era obligatorio. Al igual que en una comunidad del mundo real, fueron muchas las personas que no se lanzaron a la vida asociativa.

Sin embargo, para aquellos que querían comunidad, el registro del archivo respalda que GeoCities fue un lugar en el cual podían hallarla. La decisión consciente de desarrollar GeoCities siguiendo el modelo de vecindario ayudó a crear un sentido de comunidad virtual. La comunidad se encontraría a nivel de vecindario, desde el Heartland centrado en la familia, hasta el refugio LGBT de WestHollywood o la filosófica Athens. Por lo general, un recién llegado sería atendido por un «líder comunitario» local, otro usuario que ofrecía su tiempo –a cambio de algunos beneficios relativamente menores– para ayudar a otros a aprender diseño web básico. Los nuevos usuarios también podían solicitar a sus líderes comunitarios que reseñaran el sitio, lo que podría generar premios.

5. Encontrar una comunidad en un archivo web: Exploración computacional de GeoCities

¿Cómo podemos hallar comunidad en un archivo web? Esto es algo más fácil de decir que de hacer. Explorar actualmente las ruinas digitales de GeoCities presenta desafíos únicos. ¿Cómo se puede extraer información histórica significativa de un conjunto de datos tan masivo? Un enfoque que utilizan

los humanistas digitales para encontrar información en un cuerpo de documentos es el «modelado de temas» (*topic modelling*). Dicho en pocas palabras, es una técnica que toma una gran cantidad de texto y encuentra palabras que aparecen frecuentemente próximas entre sí (ya sea en la misma oración o cerca) y las designa como «temas» (*topics*)[13]. Cuando planteé estudiar la idea de comunidad en GeoCities, me pregunté si los temas más populares o tratados dentro de cada vecindario coincidían con lo que GeoCities preveía para ellos. Si fuera así, entonces podríamos decir que, en general, el enfoque de vecindario funcionó.

Resulta que fue así. Los temas que aparecieron en los vecindarios fueron los que en su mayoría «deberían» haber aparecido. Dos ejemplos de la cultura popular nos ayudan a entenderlo. En EnchantedForest, el área diseñada para incluir webs creados por niños y sus padres para otros niños y otros padres, tenemos un tema que consiste en parte en «pooh friends tigger winnie christopher color piglet», es decir, los personajes principales de *Winnie-the-Pooh* de A.A. Milne (Eeyore también aparece en la lista, aunque Rabbit está ausente). En Hollywood encontramos a los personajes del programa de televisión estadounidense *Friends*: «joey rachel ross monica chandler». Los barrios estaban siendo utilizados por personas de acuerdo con la previsión de GeoCities, algo que es significativo en sí mismo.

Sin embargo, no siempre ocurrió así. Un vecindario como EnchantedForest se mantuvo enfocado en los niños, debido en parte a los esfuerzos y cuidados de los líderes comunitarios ante los temores sobre la explotación infantil en línea. El vecindario del Pentágono, por otro lado, se expandió más allá de su objetivo inicial, el de ser una forma de conectar a militares desplegados por doquier y trasladados continuamente, para ser un centro de discusiones más amplias

13 Sobre el modelado de temas, véase: Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana/Chicago/Springfield: University of Illinois Press, 2013); Shawn Graham, Scott Weingart y Ian Milligan, «Getting Started with Topic Modeling and MALLET,» *Programming Historian*, (2012), http://programminghistorian.org/lessons/topic-modeling-and-mallet.html.

sobre historia militar, entrando incluso en el activismo y la discusión política. El centro de la moda, FashionAvenue, se convirtió en algo más que un lugar para debatir sobre el estilo, pues también trató sobre concursos de belleza, y fue un espacio para compartir imágenes de las estrellas favoritas de cine y televisión. Heartland anticipó una visión particular de la «familia» centrada en la fe cristiana, en cuestiones internas y en genealogía.

Si podemos leer a distancia el contenido textual a través del modelado de temas, también podemos leer a distancia el contenido web mirando muchas imágenes[14]. El análisis de la imagen nos puede ofrecer una idea de cómo funcionaban los vecindarios. Lo que se hizo fue extraer todas las imágenes de cada vecindario y se organizaron como montajes[15]. Estas composiciones nos permiten ver los contornos generales de una comunidad, aunque deben usarse con precaución[16]. Por ejemplo, al observar todas las imágenes del EnchantedForest enfocado en los niños, podemos ver rápidamente que la mayoría de ellas son personajes de dibujos animados. De un vistazo, tenemos una idea del vecindario sin necesidad de visitar cada página.

También podemos hacernos una idea de cómo las personas tomaron prestadas y adaptaron imágenes por todo el EnchantedForest. Por ejemplo, podemos descubrir con qué frecuencia se distribuyó exactamente la misma imagen o GIF animado en el vecindario de los niños. A partir de eso,

14 He tratado en profundidad el uso de imágenes en: Ian Milligan, «Learning to 'See' the Past at Scale: Exploring Web Archives through Hundreds of Thousands of Images,» en *Seeing the Past with Computers* (Ann Arbor: University of Michigan Press, 2018), 116–36.

15 La guía práctica se puede consultar en: «GUIDE TO VISUALIZING VIDEO AND IMAGE SEQUENCES,» https://docs.google.com/document/d/1PqSZmKwQwSIFrbmVi-evbStTbt7PrtsxNgC3W1oY5C4/edit.

16 Por ejemplo, las imágenes se organizan en un montaje sin relación con las demás, y los estudiosos han señalado que tendemos a privilegiar las relaciones de arriba hacia abajo sobre las relaciones de izquierda a derecha, incluso siendo idénticas. Véase: Daniel R. Montello y otros. «Testing the First Law of Cognitive Geography on Point-Display Spatializations,» en *Spatial Information Theory. Foundations of Geographic Information Science. COSIT 2003. Lecture Notes in Computer Science*, vol 2825 (2003). Springer, Berlin, Heidelberg, ed. Walter Kuhn, Michael F. Worboys y Sabine Timpf: 316–331, https://doi.org/10.1007/978-3-540-39923-0_21.

advertí que un GIF animado de Tigger, de la serie animada Winnie the Pooh, era la undécima imagen más popular en el EnchantedForest, apareciendo 48 veces. Esta sensación de préstamo y cohesión aparece en muchos vecindarios de GeoCities. Las comunidades culturales populares contienen capturas de pantalla de programas de televisión y de películas populares; Athens, por ejemplo, contiene una cantidad desproporcionada de imágenes en blanco y negro –tras examinarlo, se trata de figuras históricas, indicadoras de los fundamentos educativos de la comunidad–. Tanto en el análisis de imágenes como en el modelado de temas, las sorpresas son pocas y espaciadas. Generalmente encontramos lo que esperaríamos encontrar. Este es un descubrimiento significativo en sí mismo.

Sin embargo, con eso no basta, pues hay que leer las páginas individuales para enterarse de los nodos significativos de cada vecindario. De hecho, sin leer realmente páginas web, ¿puede un historiador entender verdaderamente un archivo web? Ahora bien, con tantas páginas, lo difícil para el historiador es determinar cuáles ha de mirar.

Un punto de partida razonable es encontrar las páginas que la mayoría de los usuarios parecen estar mirando, en función de sus patrones de hipervínculos. Los enlaces son uno de los elementos de referencia en la web. Si suponemos que el hipervínculo es una práctica deliberada, donde enlaces se usan para conectarnos con asuntos de interés, entonces podemos comenzar a usar estos enlaces para construir la web como una red social. Durante la década de 1990 y la del apogeo de GeoCities, los enlaces fueron especialmente importantes ya que los motores de búsqueda eran comparativamente rudimentarios.

Explorar hipervínculos en el Enchanted Forest da una idea de los contornos generales de la comunidad, así como de los temas principales que se encuentran en su ascenso y caída. Muchos usuarios de GeoCities querían ser descubiertos: hacer que otros usuarios encontraran su sitio y se involucraran con el contenido. Un testimonio de ello fue la ubicuidad casi

total de libros y de contadores de visitas. A fines de los '90, la inclusión de una web en los motores de búsqueda no era automática: muchos requerían rellenar un formulario si uno se quería asegurar de ser descubierto correctamente. Así pues, para encontrar determinado contenido, muchos usuarios se basaban en enlaces: de libros de visitas, que los usuarios se daban entre sí, o de anillos web (*webrings*) de los que podrían haber sido miembros.

Origen	Destino	Número de enlaces
http://geocities.com/EnchantedForest/Meadow/1134	http://www.geocities.com/EnchantedForest/1004	83
http://geocities.com/Area51/Stargate/1357	http://www.geocities.com/Area51/EnchantedForest/4213	33
http://geocities.com/Eureka/1309	http://www.geocities.com/EnchantedForest/Tower/7555	27

Tabla 1: Enlaces dentro de EnchantedForest
Fuente: elaborado por el autor.

En consecuencia, se extrajeron todos los enlaces del Enchanted Forest. La Tabla 1 muestra un ejemplo de las relaciones enlazadas encontradas. La tabla demuestra que en todas las páginas de EnchantedForest/Meadow/1134 –el Enchanted Forest Meadow Community Center (incluidas numerosas subpáginas, como el boletín Meadow, el anillo web, el registro de líderes de la comunidad, etc.)– había 83 enlaces al sitio EnchantedForest/1004: el principal centro comunitario. Los «centros comunitarios» eran centros primarios de actividad dentro de GeoCities, ya que tenían como objetivo reunir a un vecindario: un lugar para compartir información sobre sitios, destacar páginas particularmente exitosas, promover «premios», tal vez llevar un periódico para el debate; en resumen, un lugar para que los usuarios se reunieran dentro de espacios concretos de GeoCities. En el ejemplo de Meadow, fue un caso de comunidad para el suburbio de «Meadow»: los pocos miles de sitios que conformaban todas las páginas de esa parte de GeoCities. La tabla anterior muestra cómo los

centros comunitarios se comunicaban entre sí: el principal de todo el Enchanted Forest, que vinculaba en gran medida a sus ramificaciones suburbanas.

Si proyectamos eso mismo a todo GeoCities, entonces podamos comenzar a ver con qué webs de Enchanted Forest se enlazaba más, cuáles eran aquellos que tenían más enlaces hacia otros y qué sitios eran aquellos con los que era más probable que uno se hubiera tropezado al aventurarse aleatoriamente a través del sitio. Para encontrar páginas significativas, se utilizó el algoritmo PageRank. PageRank está actualmente en el centro del motor de búsqueda de Google y, de manera similar, puede ayudarnos a encontrar sitios útiles. Explicado brevemente, PageRank se basa en los enlaces existentes a determinados sitios, cada uno de los cuales puede considerarse un «voto de confianza» a ese sitio; sin embargo, estos votos se ponderan de acuerdo con el PageRank del sitio que enlaza, lo que nos ayuda a solucionar el problema de los sitios que ofrecen muchos enlaces pero que no tienen valor en sí mismos[17].

17 Sepander Kamva y otros, «Exploiting the Block Structure of the Web for Computing PageRank» (Stanford, 2003), http://ilpubs.stanford.edu:8090/579/1/2003-17.pdf; «Pagerank Explained Correctly with Examples, de Ian Rogers,» *Page Rank Explained*, acceso el 14 de abril de 2020, http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm.

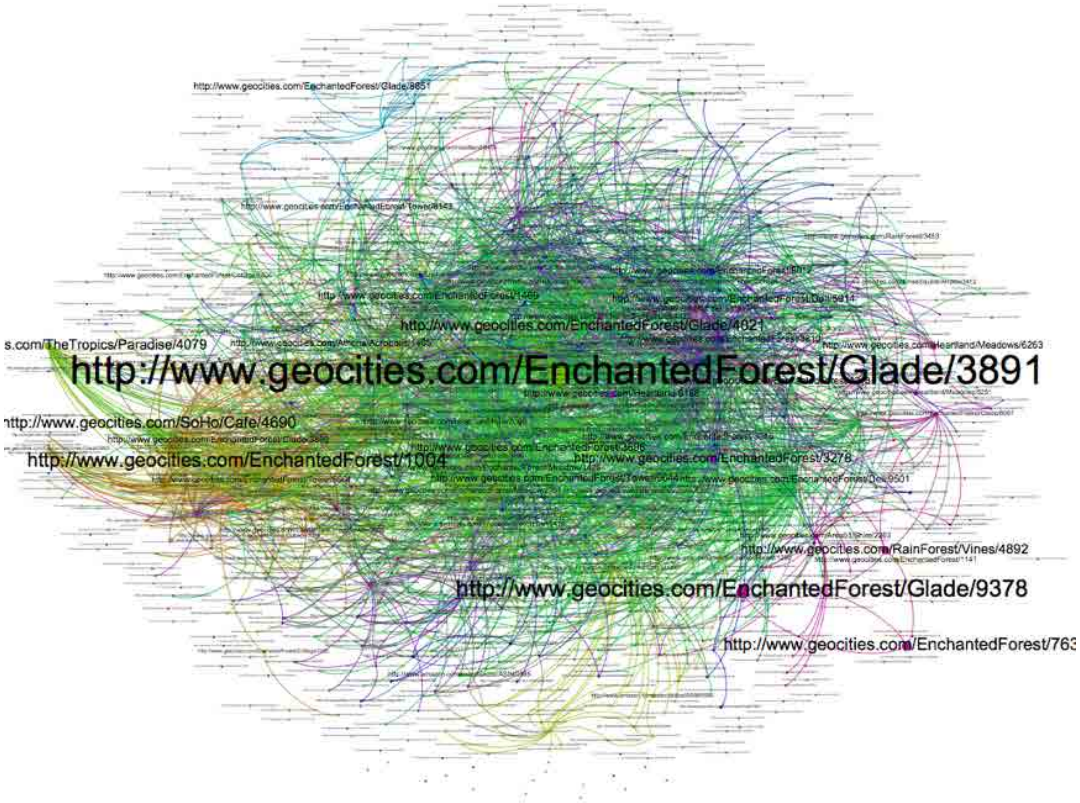

Figura 1: Visualizando la Estructura de Enlaces del EnchantedForest

El Enchanted Forest era una comunidad interconectada, como se ve en la Figura 1. Si bien las tablas, como la Tabla 1, pueden ayudarnos a dar sentido a los sitios individuales, también puede ser útil para ver las estructuras de hipervínculos como redes. La figura muestra todos los sitios del Enchanted Forest y dibuja líneas para cada uno de los hipervínculos

existentes entre ellos. Una «bola de pelos» (*hairball*) como la de la figura demuestra cuán interconectada estaba la red: los sitios enlazaban mucho entre sí, y varios de ellos estaban conectados a muchos otros. Los nodos (webs) se dimensionan de acuerdo con su PageRank, y arriba podemos ver algunos sitios con altas puntuaciones de PageRank. Esto nos ayuda a explorar algunos de los sitios más populares. Estos iban desde aquellos destinados a los niños, los que construyeron comunidad a través de la provisión de servicios o de premios y, por supuesto, aquellos escritos por los propios niños. Lo que generalmente compartieron fue una estrecha conexión con el ecosistema GeoCities: los sitios habían recibido numerosos galardones, compartieron imágenes con otros, se convirtieron en páginas destacadas y, por lo general, se implicaron en la comunidad vibrante que se estaba formando.

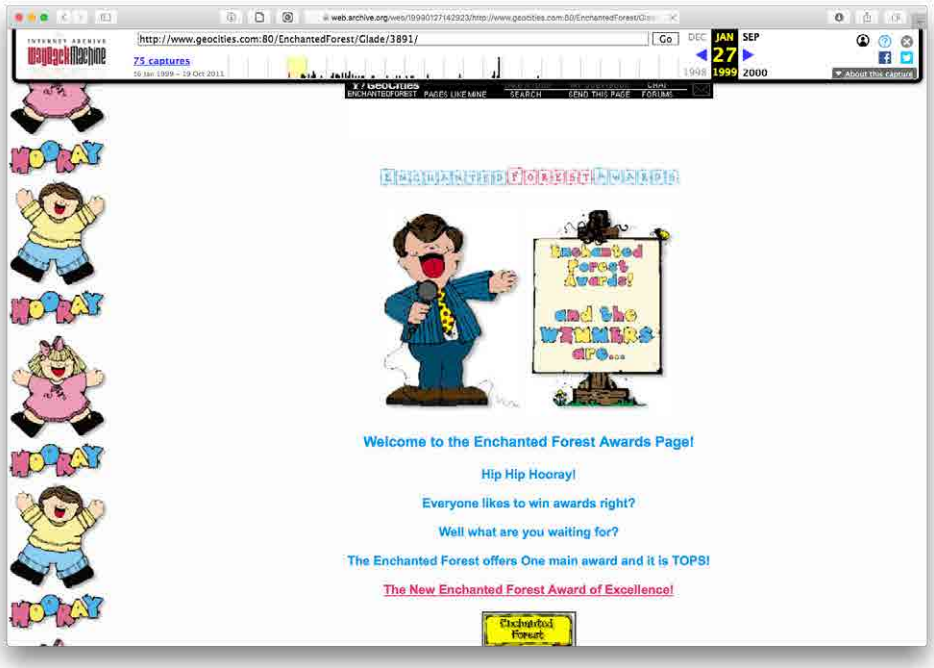

Figura 2: las páginas mejor clasificadas en el Enchanted Forest

Fuente: «EnchantedForest Award,» Internet Archive Waybackmachine, http://web.archive.org/web/20010721183753/http://www.geocities.com/EnchantedForest/Glade/3891/.

Algunos de los sitios más destacados desempeñaban funciones conectivas dentro de GeoCities: muchos sitios enlazaban con ellos, y a su vez estos enlazaban a muchos otros. El sitio con el PageRank más alto, EnchantedForest/Glade/3891 (Figura 2), fue la Enchanted Forest Awards Page[18]. Regía la concesión del premio principal distribuido en el vecindario, el Enchanted Forest Award of Excellence. Glade/3891 era una web semioficial, operada por líderes comunitarios voluntarios pero con el fuerte respaldo de los administradores de GeoCities, que mantenía un procedimiento de solicitud para aquellos que querían obtener el premio. En la práctica, si estos sitios llegaban a ser exitosos y se les daba un símbolo, o insignia, entonces podían alojar a otros en sus propias páginas. Cada una de estas insignias, por supuesto, contenía un enlace de retorno al adjudicatario, creando muchos hipervínculos, que luego aparecían en el análisis de PageRank.

La combinación de modelado de temas, análisis de imágenes y análisis de redes sugiere varios hallazgos. Primero, muestra que una combinación de enfoques computacionales y de lectura de las páginas identificadas por esos algoritmos puede proporcionar una vía creíble para la historia de la web. En segundo lugar, demuestra que, en líneas generales, el sistema de vecindario de GeoCities funcionó: el vecindario de los niños se preocupaba por los niños, Heartland se preocupaba por temas familiares (aunque más estrechamente definidos) como la genealogía y los valores cristianos, y en el Pentagon dominaron los temas militares. Y, finalmente, demuestra que tales cosas no ocurrieron por casualidad. En el Enchanted Forest, por ejemplo, los sitios mejor clasificados según PageRank tenían que ver con la comunidad: «líderes comunitarios», premios otorgados. En resumen, la comunidad parece haber sido promulgada deliberada y físicamente; sin embargo, repasar estos sitios comienza a plantear cuestiones éticas: pasemos ahora a ellas.

18 Se puede visitar en: «EnchantedForest Award,» Internet Archive Waybackmachine, http://web.archive.org/web/20010721183753, /http://www.geocities.com/EnchantedForest/Glade/3891/.

6. La vida de los otros: desafíos éticos con los archivos web

Trabajar con contenido creado por millones de personas comunes requiere cuidado[19]. En lo que a esto respecta, sus sitios son diferentes a los de las organizaciones públicas, como partidos políticos o corporaciones. Dichas entidades públicas no tienen expectativas razonables de privacidad. Pero, ¿qué sucede cuando dirigimos nuestra atención a la gente común que ha publicado contenido en línea? La línea entre una página web pública y otra privada no siempre es clara, por supuesto. Un bloguero que escribe para sitios prominentes, pero tiene menos de 20.000 seguidores en Twitter, ¿es una figura pública? El Rector de una universidad casi seguro que sí es una figura pública, pero ¿lo es un profesor?, ¿y un estudiante graduado? En todos los casos, el contexto importa. Trabajar con archivos web requiere navegar constantemente por las zonas grises[20].

Las cosas parecen más claras en cualquiera de los extremos de la expectativa del espectro de privacidad. Si bien, es casi seguro que un político no tiene tales expectativas, podemos asegurar lo contrario de un niño de catorce años que escribió su página web en 1996. Incluso la generación actual de jóvenes, que han crecido en un entorno mediático y digital, tiene un sorprendente grado de ingenuidad acerca de cuán accesible puede ser el material que publican[21]. No es razonable

19 Una meditada visión de conjunto en: «Mining Social Media Data: How Are Research Sponsors and Researchers Addressing the Ethical Challenges?,» *Research Ethics* Vol. 14, nº 2 (2017): 1-39, DOI: https://doi.org/10.1177/1747016117738559. attitudes, feelings and relationships are increasingly being harvested from social media platforms and re-used for research purposes. This can be ethically problematic, even where such data exist in the public domain. We set out to explore how the academic community is addressing these challenges by analysing a national corpus of research ethics guidelines and published studies in one interdisciplinary research area. Methods: Ethics guidelines published by Research Councils UK (RCUK).

20 Ella Dawson, «The Dark Side of Going Viral», *Vox*, Washington/Nueva York, 10 de julio de 2018, https://www.vox.com/first-person/2018/7/10/17553796/plane-bae-viral-airplane-romance.

21 Una encuesta encargada por *The Atlantic* y el Instituto Aspen, por ejemplo, encontró que «los estadounidenses más jóvenes expresan una mayor expectativa de que la información de la persona que usan en sitios como Facebook y Twitter será privada. Un poco más de la mitad de los jóvenes de 18 a 29 años dijeron que tenían estas

esperar que, en buena medida, los archivistas de la web tengan que despejar estas dudas. El Internet Archive, por ejemplo, simplemente no tiene los recursos necesarios para tomar una determinación de tipo ético página a página. Más bien, la responsabilidad ética recae en los investigadores. Si bien rara vez hay respuestas éticas directas, los historiadores que trabajan con archivos web deben considerar significativamente cuestiones clave como las de consentimiento y perjuicio.

Nada de esto será sencillo. La naturaleza algorítmica del archivo web supone que un historiador no necesariamente pueda recurrir a un archivista web para obtener orientación. Los archivistas web generalmente no tienen comunicación directa con los donantes de material archivado, y en cualquier caso las juntas de ética institucional podrían desaprobar que los archivistas o investigadores contactasen con los autores de una web.

Las preguntas comienzan a aparecer casi de inmediato cuando se trabaja con colecciones como GeoCities. ¿Qué pasa si te encuentras con páginas donde la gente está reflexionando abiertamente acerca del suicidio? ¿Y si están expresando un dolor muy personal por la pérdida de un hijo, o escribiendo sobre sus dudas y luchas con respecto a ser padres? O, en una nota menos taciturna, ¿qué ocurre si están celebrando haber contactado con una vibrante comunidad gay y lesbiana en línea mientras están aprisionados en su «vida real» en un país del sudeste asiático? Si bien en las últimas dos décadas han ido evolucionando las percepciones que los usuarios tienen de la privacidad, resulta muy fácil encontrar información sobre las personas. Recordemos el contexto de lo pobre que era buscar en la web en la década de 1990. Algo de esta información, por otro lado, podría haber sido compartido en el contexto de un «público íntimo» –partes de la web que operan con una naturaleza limitada, a pequeña escala–, teóricamente

expectativas, mientras que solo el 38 por ciento de los estadounidenses mayores de 65 años dijeron lo mismo». Para profundizar en ello: Rebecca J. Rosen, «59% of Young People Say the Internet Is Shaping Who They Are», *The Atlantic*, Washington D.C., 27 de junio de 2012, https://www.theatlantic.com/technology/archive/2012/06/59-of-young-people-say-the-internet-is-shaping-who-they-are/259022/.

accesible para cualquiera con un navegador, pero caracterizada por tener unos lazos emocionales muy estrechos[22].

No hay respuestas directas a ninguna de las preguntas anteriores, que se basan en dilemas reales que los estudiantes o yo hemos encontrado. Por un lado, los archivos web documentan historias fascinantes que arrojan luz sobre la cultura y la actividad humana; por otro lado, contienen información personal que tal vez los afectados no quieran asociar con sus nombres cuando pasaron más de veinte años desde que publicaron el material en la red. Dada la importancia de las historias de la gente común, por supuesto, no resulta ético *no* recopilar estos relatos. Son importantes contrapesos a las historias de los poderosos y dominantes.

El reciente trabajo realizado por Sarah McTavish, una graduada que está terminando su doctorado en la Universidad de Waterloo, propició que comprendiera los problemas éticos a los que nos podemos enfrentar en la investigación histórica basada en la web. McTavish explora la identidad gay y lésbica en línea. En el vecindario de GeoCities de West Hollywood, encontró docenas de webs de personas LGBT: completos en algunos casos, con nombres y apellidos, con información personal, con información de contacto y mucho más[23]. En su caso, ha de lidiar con preguntas muy apremiantes. ¿Qué debe mostrar en las presentaciones que emplea para dar una conferencia, y qué debe desdibujar? ¿Qué nombres debe proporcionar y cuáles no? Y, cuando se trata de publicaciones y de la tesis, ¿qué forma deben tener las citas? ¿Es una investigación histórica «auténtica» si las citas no apuntan al documento histórico exacto?

Afortunadamente, hay una comunidad de académicos que se ha enfrentado a estas preguntas: académicos de la «web en vivo» (*live web*). Las prácticas actuales de los

22 Aimée Morrison, «'Suffused by Feeling and Affect': The Intimate Public of Personal Mommy Blogging,» *Biography,* nº 34.1 (2011): 37–55.

23 Sarah McTavish, «West Hollywood Goes Global: Exploring Global Queer Identity in GeoCities» (presentación en *Global Digital Humanities Symposium*, Michigan State University, 22-23 de marzo de 2018).

investigadores de los Estudios de Internet se definen por pautas y consideraciones establecidas por asociaciones académicas e investigadores líderes en tal campo, más que por regulaciones prescriptivas. Generalmente operan bajo el supuesto de que los materiales de acceso abierto basados en la web son publicaciones, pero no creen que eso les dé carta blanca para usar el material. Aunque es legal citar tuits, blogs o webs, eso no necesariamente lo hace ético. Como ha señalado Aaron Bady, un reflexivo bloguero, periodista, académico y comentarista, «el acto de vincular o citar a alguien que no considera su Twitter como público solo es éticamente correcto si consideramos que la ley está por encima de la ética del consentimiento»[24]. El tecnólogo Anil Dash también se apresura a aclarar la «legitimidad» del uso de estas fuentes: «¿Cuándo acordamos permitir que los medios redefinan como uso legítimo cualquier uso de las redes sociales, sin recurso al consentimiento ni marco en el que situarlo?»[25]. Desde el campo de los estudios de la información, Safiya Noble argumenta que «hay muchos aspectos problemáticos como para que nos satisfaga el que cada una de nuestras acciones en el registro digital se conserve permanentemente o que se guarde durante tanto tiempo como para que tenga un impacto duradero en nuestras vidas personales»[26]. Si bien su énfasis particular está en las prácticas de Google, la cuestión se aplica a las redes sociales o la investigación en archivos web.

Las recientes tendencias en el pensamiento archivístico también han ayudado a cuestionar los supuestos implícitos en torno a la idea de «apertura», que es omnipresente en los archivos occidentales y en la investigación basada en la web en general. Mi propio sesgo es hacia la apertura siempre que sea posible: una visión del mundo influida por las normas occidentales en torno a la investigación académica y a la

24 Aaron Bady, «#NotAllPublic, Heartburn, Twitter», *The New Inquiry*, Nueva York, 10 de junio de 2014, https://thenewinquiry.com/blog/notallpublic-heartburn-twitter/.

25 «What Is Public? It's so Simple, Right?, de Anil Dash,» *Medium*, 24 de julio de 2014, https://medium.com/message/what-is-public-f33b16d780f9

26 Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (Nueva York: NYU Press, 2018), 129.

información; sin embargo, Kimberly Christen ha subrayado la importancia de los grupos –en su caso, las comunidades indígenas de todo el mundo– a la hora de dictar sus propias restricciones de acceso «basadas en parámetros culturales»[27]. Christen y sus colaboradores han puesto en práctica la teoría con el Mukurtu Content Management System, que proporciona restricciones de acceso granular al contenido. Ella escribe:

En el caso de la comunidad Mukurtu, el Mukurtu Wumpurrarni-kari Archive, el acceso a ciertos materiales culturales (y el conocimiento que anima a estos materiales) se decide según un sistema dinámico de responsabilidad donde la edad, el género, el estatus ritual, la familia y todas las relaciones basadas en el lugar se combinan (y se recombinan a medida que las afiliaciones cambian a lo largo de la vida) para producir un acceso continuado a los materiales dentro de la comunidad[28].

Si bien esto parece antitético a la dualidad cerrado/abierto de los métodos tradicionales de acceso, estos enfoques ayudan a generar debate y plantean preguntas sobre cuál puede ser el enfoque correcto para el acceso al patrimonio cultural. Los protocolos culturales deben estar en el centro de las preguntas de acceso, ya sea discutiendo sobre comunidades indígenas o sobre subculturas de internet muy entrelazadas; sin embargo, la escala es, como hemos visto, el factor que lo complica todo. El modelo granular de Mukurtu cuestiona admirablemente la dualidad de archivos abiertos/cerrados y nos reta a reconsiderar la primacía de los modelos de acceso occidentales, no se adapta a los miles de millones de recursos encontrados en un archivo web. Como hemos visto, los creadores y propietarios de contenido son difíciles de identificar y, a menudo, es imposible contactar con ellos. Esto coloca a los recopiladores en un dilema difícil, ya que se verían obligados a tomar decisiones de acceso en nombre de terceros.

27 Kimberly Christen, «Opening Archives: Respectful Repatriation,» *The American Archivist* Vol. 74, n° 1 (2011): 186, https://doi.org/10.17723/aarc.74.1.4233nv6nv6428521.

28 Christen, «Opening Archives,» 189.

En consecuencia, con los archivos web, la responsabilidad ética a menudo recae en los propios investigadores.

Esta responsabilidad investigadora se basa en los tres supuestos principales de muchos recopiladores e investigadores: que la web es un medio de publicación, que la «naturaleza efímera [de la web] es una deficiencia del medio» que pueden «reparar» los archivistas de la web y personal de archivo capacitado, y que los archivos web están recolectando «material que está disponible de todos modos»[29]. Como señala el archivista Eira Tansey, «la idea que, si el contenido personal auto-publicado se puede encontrar públicamente en la web, y que entonces se considera uso legítimo (*fair game*) su reutilización periodística, académica o de archivado, es tan común que pocos la cuestionan»[30]. A medida que utilizamos archivos web, debemos preguntarnos sobre cuestiones éticas y de privacidad, especialmente dada la escala, el alcance y la invisibilidad de muchos esfuerzos de archivado web.

Por supuesto, como historiadores también debemos darnos cuenta de que el contexto histórico en el que se crearon las fuentes nos informa cómo podemos usarlas éticamente en la actualidad. El ejemplo LGBT de McTavish ha mostrado que los usuarios pusieron enormes cantidades de información personal en línea en GeoCities, a menudo cosas que no querrían que la gente supiera en el «mundo real»: sus preferencias sexuales, direcciones y demás. La gente también hacía bromas o decía cosas en línea sin darse cuenta de las implicaciones de tener un registro de acceso público que se mantiene durante años, con normas sociales que han ido cambiando, con la celebridad personal en auge o decayendo, y otros aspectos. Como Megan McArdle ha señalado en el Washington Post.

29 Andreas Rauber, Max Kaiser, and Bernhard Wachter, «Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda,» *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.

30 «My Talk at Personal Digital Archiving 2015, de Eira Tansey,» *Eiratansey.Com* (blog), 16 de agosto de 2015, http://eiratansey.com/2015/08/16/my-talk-at-personal-digital-archiving-2015/.

A menudo, las tecnologías emergentes de comunicación se utilizan inicialmente de manera informal, de modo que los primeros usuarios las ven como una especie de espacio privado ampliado... deberíamos mostrar cierta consideración con aquellos usuarios iniciales que no se dieron cuenta de que los blogs o las redes sociales terminarían siendo más como un periódico que como el taburete de un bar[31].

Todos estos aspectos sugieren que los historiadores tienen la obligación de cuidar sus fuentes históricas y a sus individuos. Los historiadores necesitan comprender el contexto, los protocolos culturales y las expectativas de privacidad cuando usan estos recursos. La escala puede mitigar, pero no eliminar los problemas; del mismo modo podemos cambiar nuestros patrones de citas o hacer anónimas estas fuentes impresas. Puede que las citas no estén allí, pero podemos confiar unos en otros en que estamos haciendo un trabajo histórico bueno y responsable.

Para finalizar, ¿cómo puede un investigador usar éticamente un gran archivo web, como GeoCities? Primero, podemos usar la «lectura distante» para alejar nuestra mirada de los sitios web individuales y buscar patrones más grandes dentro de un archivo. Hemos visto algo de ello en este artículo: mirando patrones de enlaces para confirmar el PageRank, explorando temas, mirando miles de imágenes en lugar de decenas. Nada de esto elimina, de ninguna manera, todas las preocupaciones éticas en torno a una recopilación –pensemos en lo que la NSA ha hecho con tipos similares de datos– pero sí los mitiga hasta cierto punto. Las personas tienen protegida su privacidad, y otros no pueden encontrar sus sitios si no tienen el mismo nivel de acceso al archivo. Las personas están ocultas, pero todavía se las puede leer en el registro histórico. También nos habla de un método de investigación valioso, ya que, por muy diligente o completo

31 Megan McArdle, «People Are Getting Fired for Old Bad Tweets. Here's How to Fix It», *Washington Post*, Washington, 24 de julio de 2018, https://www.washingtonpost.com/opinions/we-need-a-statute-of-limitations-on-bad-tweets/2018/07/24/a84e335c-8f7d-11e8-b769-e3fff17f0689_story.html.

que sea un investigador, no podrán leer todas las páginas de GeoCities.

Y sin embargo, y a la postre, los historiadores aún necesitan leer documentos individuales, o al menos exponer claramente por qué no lo hacen, si finalmente oscurecen la fuente. Necesitan citar, poner notas al pie remitiendo a las fuentes con enlaces. Las normas profesionales exigen una investigación verificable y comprobable. Pero las cosas son complicadas en el caso de los archivos web, cuando un documento está a un clic de distancia y ya no hay que tomar un avión para obtener un archivo remoto.

En última instancia, y de manera imperfecta, deberá recaer en los investigadores individuales la responsabilidad de llevar a cabo una evaluación de riesgos y considerar el contexto en el que se publicó el material que están leyendo. El autor de la web individual que están citando o utilizando como evidencia, ¿tenía una expectativa razonable de privacidad? Si fuera un líder comunitario de GeoCities, con enlaces entrantes de cientos de otros sitios y que aparecía en directorios, probablemente, no. Si estaba publicando un relato sincero en el libro de visitas de GeoCities de un amigo, parte de una red social aparentemente cerrada de algunos compañeros de la escuela secundaria, probablemente, sí. Esto significa que, al interactuar con sitios individuales, la medida fundamental debe ser la expectativa de privacidad. Una web con enlaces a muchas otras webs, con un orgulloso y prominente contador en el margen que muestra los miles de visitantes que entraron, indica que es una web cuyo propietario quería ser leído y encontrado. Una web más pequeña, sin contador de enlaces, dirigida principalmente a amigos, escrita por un adolescente con mensajes e imágenes reveladores, supone una fuente éticamente más dudosa. Al menos, necesitamos pararnos y pensar en nuestras obligaciones como historiadores.

Debemos utilizar fuentes como GeoCities con moderado entusiasmo. Tenemos poder porque podemos acceder a los blogs, reflexiones y momentos personales de literalmente

millones de personas a las que nunca antes se habría accedido, pero necesitamos usar este poder de manera responsable.

Conclusiones

Nuestra sociedad solía olvidar –dicho de otra manera, no solíamos dejar tantos rastros de nosotros mismos–. Ahora podemos recordar mejor, pero en una escala que cambiará decisivamente la forma en que trabajan los historiadores. Tradicionalmente, el registro histórico estaba sesgado, en favor de aquellos que, debido a sus posiciones de privilegio e influencia, han podido inscribirse en el registro histórico, así como de aquellos que se encontraron allí por razones de descrédito. La web cambia todo eso. Esto no significa que el registro histórico de la web sea democráticamente representativo de la sociedad (todavía existen barreras considerables para el acceso y la publicación en la web en función de la raza, el origen étnico, la clase y el género), la web brinda a más personas que nunca la oportunidad de estar en el registro histórico. Ahora se registran cosas que nunca antes se habrían registrado, por personas que nunca antes habrían dejado una vida documentada para los historiadores.

Este artículo ha argumentado que este es un gran cambio y que merece nuestra atención. Los estudios que cubran períodos posteriores a 1996 –año en que el archivado de la web comenzó a generalizarse con el Internet Archive con sede en San Francisco y en varias bibliotecas nacionales de todo el mundo– solo serán creíbles si incorporan esta información digital nativa. Imaginemos escribir una historia de la presidencia de Donald Trump, de la pandemia global del COVID-19 o de los ataques terroristas del 11 de septiembre sin usar webs archivadas. O imaginemos abordar el tema de la guerra de Irak sin considerar las publicaciones y pensamientos que los soldados que estuvieron en el campo de batalla expusieron por toda la web. Lo mismo ocurre con cualquier cantidad de temas sociales y culturales, desde celebridades como Michael Jackson hasta fenómenos políticos y sociales como la lucha contra el cambio climático. Sería intelectualmente deshonesto abordar esos temas sin recurrir a la web.

Los historiadores deben estar preparados para afrontar las fuerzas a partir de lo que se ha esbozado a lo largo de este artículo. Hasta ahora, he discutido principalmente sobre historiadores en abstracto –como miembros de una profesión dedicada a investigar y escribir sobre el pasado–. Pero los historiadores también constituyen una profesión en virtud de su formación y estándares, sociedades, redes profesionales y cultura. Este último factor, la cultura, es quizás el obstáculo más importante que se interpone en el camino hacia el uso histórico significativo de los archivos web. Los historiadores pueden tener computadoras potentes, disponibles para analizar archivos web completos, pero deben estar dispuestos y ser capaces de estar a la altura de las circunstancias. Para que este tipo de cambio se materialice debe haber incentivos. ¿Cómo es esto?

Ocurre que la profesión sigue dedicada en gran medida a la erudición textual tradicional. Podemos ver este atrincheramiento en la renuencia profesional a participar en dos dimensiones clave del trabajo digital: modelos colaborativos de investigación y análisis cuantitativo. El historiador James Baker ha señalado que, si bien, es necesario comprender las metodologías cuantitativas para que prospere la historia digital, los libros de texto recientes sobre métodos históricos estándar han eliminado las secciones cuantitativas[32]. Más allá de los libros de texto, también hay luchas por el reconocimiento de la investigación digital dentro de los departamentos de historia. Los académicos capacitados digitalmente no son parias –son contratados y realizan investigaciones de alto perfil–, pero los estándares de promoción y estabilidad en muchas universidades de investigación continúan insistiendo en las monografías tradicionales para el progreso profesional. A medida que la investigación cambia de forma, también lo hace nuestra profesión.

El camino que nos queda por delante no será sencillo, pero vale la pena recorrerlo. Se requerirá que los historiadores

32 «Digital History and the Death of Quant, de James Baker,» *British Library Digital Scholarship Blog*, 5 de abril de 2014, http://britishlibrary.typepad.co.uk/digital-scholarship/2014/04/digital-history-and-the-death-of-quant.html.

cambien: adoptar nuevos estándares y prácticas en nuestro oficio, desarrollar nuestro conocimiento sobre computadoras y algoritmos, y trabajar de manera más colaborativa, formando equipos con colegas en humanidades digitales y ciencias de la computación. Requerirá una mayor participación ética, en pensamiento y práctica. Implicará cometer errores en el camino, a medida que los historiadores encuentren su camino por este nuevo territorio. En definitiva, valdrá la pena. Los archivos web ofrecen la posibilidad de incorporar más voces y más personas. Una historia más inclusiva está a la vuelta de la esquina. Necesitamos estar preparados.

Bibliografía

Christen, Kimberly. «Opening Archives: Respectful Repatriation.» *The American Archivist* Vol. 74, n° 1 (2011): 185-210. DOI: https://doi.org/10.17723/aarc.74.1.4233nv6nv6428521.

Graham, Shawn, Scott Weingart, e Ian Milligan. «Getting Started with Topic Modeling and MALLET.» *Programming Historian*, (2012). http://programminghistorian.org/lessons/topic-modeling-and-mallet.html.

Graham, Shawn, Ian Milligan y Scott Weingart. *Exploring Big Historical Data: The Historian's Macroscope*. Londres: Imperial College Press, 2015.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Urbana/Chicago/Springfield: University of Illinois Press, 2013.

Kamvar, Sepander y otros, «Exploiting the Block Structure of the Web for Computing PageRank» (Stanford, 2003), http://ilpubs.stanford.edu:8090/579/1/2003-17.pdf

McTavish, Sarah. «West Hollywood Goes Global: Exploring Global Queer Identity in GeoCities.» Presentación a *Global Digital Humanities Symposium*, Michigan State University, 22-23 de marzo de 2018.

Milligan, Ian. «Welcome to the Web: The Online Community of GeoCities and the Early Years of the World Wide Web.» En *The Web as History*, editado por Niels Brügger y Ralph Schroeder. Londres: UCL Press, 2017.

Milligan, Ian. «Exploring Web Archives in the Age of Abundance: The Case of GeoCities.» En *SAGE Handbook of Web History*, editado por Niels Brügger e Ian Milligan. Londres: SAGE, 2018.

Milligan, Ian. «Learning to 'See' the Past at Scale: Exploring Web Archives through Hundreds of Thousands of Images.» En *Seeing the Past with Computers*. Ann Arbor: University of Michigan Press, 2018.

Milligan, Ian. *History in the Age of Abundance? How the Web is Transforming Historical Research*. Montreal & Kingston: McGill-Queen's University Press, 2019.

Montello, Daniel R. y otros. «Testing the First Law of Cognitive Geography on Point-Display Spatializations.» En *Spatial Information Theory. Foundations of Geographic Information Science. COSIT 2003. Lecture Notes in Computer Science*, vol 2825 (2003). Springer, Berlin, Heidelberg, editado por Walter Kuhn, Michael F. Worboys y Sabine Timpf, 316–331. https://doi.org/10.1007/978-3-540-39923-0_21

Morrison, Aimée. «Suffused by Feeling and Affect': The Intimate Public of Personal Mommy Blogging.» *Biography*, nº 34. 1 (2011): 37–55.

Moschovitis, Christos J. P. *History of the Internet: A Chronology, 1843 to the Present*. Santa Barbara, Calif.: ABC-CLIO, 1999.

Motavalli, John. *Bamboozled at the Revolution: How Big Media Lost Billions in the Battle for the Internet*. Nueva York: Penguin Group, 2004.

Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. Nueva York: NYU Press, 2018.

Putnam, Lara. «The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast.» *The American Historical Review*, 121-2 (2016): 377–402.

Rauber, Andreas, Max Kaiser, y Bernhard Wachter. «Ethical Issues in Web Archive Creation and Usage–Towards a Research Agenda.» *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.

Rheingold, Howard. *The Virtual Community: Homesteading on the Electronic Frontier*. Cambridge, Mass.: MIT Press, 2000 (1993). Edición en digital. http://www.rheingold.com/vc/book/intro.html (trad. cast.: *La comunidad virtual: una sociedad sin fronteras*. Barcelona: Gedisa, 1996).

Rockwell, Geoffrey, y Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: The MIT Press, 2016.

Taylor, Joanna, y Claudia Pagliari,. «Mining Social Media Data: How Are Research Sponsors and Researchers Addressing the Ethical Challenges?.» *Research Ethics* Vol. 14, n° 2 (2017): 1-39. DOI: https://doi.org/10.1177/1747016117738559.

Vaidhyanathan, Siva. *The Googlization of Everything (And Why We Should Worry)*. Berkeley: University of California Press, 2011. (trad. cast.: *La Googlización de todo (y por qué deberíamos preocuparnos)*. México, D.F.: Editorial Océano, 2012).

Sitios web

British Library Digital Scholarship Blog. «Digital History and the Death of Quant, de James Baker.» 5 de abril de 2014. http://britishlibrary.typepad.co.uk/digital-scholarship/2014/04/digital-history-and-the-death-of-quant.html.

Eiratansey.Com (blog). «My Talk at Personal Digital Archiving 2015, de Eira Tansey.» 16 de agosto de2015. http://

eiratansey.com/2015/08/16/my-talk-at-personal-digital-archiving-2015/.

«GUIDE TO VISUALIZING VIDEO AND IMAGE SEQUENCES.» https://docs.google.com/ document /d/1PqSZmKwQwSIFrbmVi-evbStTbt7PrtsxNgC3W1oY5C4/edit.

Internet Archive Waybackmachine. «EnchantedForest Award.» http://web.archive.org/web/20010721183753/http://www.geocities.com/EnchantedForest/Glade/3891/.

Medium. «What Is Public? It's so Simple, Right? de Anil Dash.» 24 de julio de 2014. https://medium.com/message/what-is-public-f33b16d780f9

Page Rank Explained. «Pagerank Explained Correctly with Examples, de Ian Rogers.» Acceso el 14 de abril de 2020. http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm.

Periódicos

«Beverly Hills Internet, Builder of Interactive Cyber Cities, Launches 4 More Virtual Commu-nities Linked to Real Places». *Business Wire*, 5 de julio de 1995. https://web.archive.org/web/20081211170054/http://www.allbusiness.com/marketing-advertising/marketing-advertising/7191644-1.html.

Bady, Aaron. «#NotAllPublic, Heartburn, Twitter». *The New Inquiry*, Nueva York, 10 de junio de 2014. https://thenewinquiry.com/blog/notallpublic-heartburn-twitter/.

Dawson, Ella. «The Dark Side of Going Viral». *Vox*, Washington /Nueva York, 10 de julio de 2018, https://www.vox.com/first-person/2018/7/10/17553796/plane-bae-viral-airplane-romance.

Fletcher, Dan. «Internet Atrocity! GeoCities' Demise Erases Web History». *Time*, Nueva York, 9 de noviembre de 2009. http://

content.time.com/time/business/article/0,8599,1936645,00.html.

McArdle, Megan. «People Are Getting Fired for Old Bad Tweets. Here's How to Fix It». *Washington Post*, Washington, 24 de julio de 2018. https://www.washingtonpost.com/opinions/we-need-a-statute-of-limitations-on-bad-tweets/2018/07/24/a84e335c-8f7d-11e8-b769-e3fff17f0689_story.html.

Rosen, Rebecca J. «59% of Young People Say the Internet Is Shaping Who They Are». *The Atlantic*, Washington D.C., 27 de junio de 2012. https://www.theatlantic.com/technology/archive/2012/06/59-of-young-people-say-the-internet-is-shaping-who-they-are/259022/.

Citar este artículo
Milligan, Ian. «La historia en la era de la abundancia: archivos web e investigación histórica.» *Historia Y MEMORIA*, n° Especial (2020): 235-269. DOI: https:/doi.org/10.19053/20275137.nespecial.2020.11587.