Escola Superior d'Arxivística i Gestió de Documents
Adscrita a la UAB

EL *WEB ARCHIVING* Y LA ARCHIVÍSTICA

SANTI LOPERA LOPERA

Director del trabajo: Joan Soler

Año de elaboración: 2013

Trabajo de investigación del Máster de Archivística y Gestión de Documentos

Escola Superior d'Arxivística i Gestió de Documents

Col·lecció Treballs fi de màster i de postgrau

Cómo citar este artículo: Lopera Lopera, Santi. (2013) *El web archiving y la archivística*. Trabajo de investigación del Máster de Archivística y Gestión de Documentos de la Escola Superior d'Arxivística i Gestió de Documents. (Treballs fi de Màster i de postgrau). Http://... (consultat el ...)

Resumen

El archivo de las páginas web es una disciplina que tiene su origen en el campo de la biblioteconomía y de las ciencias de la información y es ajena al mundo archivístico de nuestro país. La primera parte del presente trabajo ofrece un breve estado de la cuestión sobre el archivo de las páginas web y, desde una perspectiva archivística, intentará dar respuestas a preguntas como ¿en qué consiste el archivo de las páginas web? ¿Para qué sirve? ¿Desde cuándo se practica? ¿Qué organizaciones lo llevan a cabo? ¿Cómo se captura y almacena la web? En la segunda parte se propone una reflexión sobre la aplicación del archivo web desde la disciplina archivística.

Palabras clave: Preservación digital, archivo web, archivística, Internet, Bibliotecas Nacionales, documentos electrónicos, tecnologías de la información y la comunicación.

Títol: El *web archiving* i l'arxivística

Resumen

L'arxivament del web és una disciplina que te el seu origen en el camp de la biblioteconomia i les ciències de la informació i és aliena al món arxivístic del nostre país. La primera part del present treball ofereix un breu estat de la qüestió sobre l'arxivament de les pàgines web i, des d'una perspectiva arxivística, intentarà donar resposta a qüestions com en què consisteix l'arxivament de les pàgines web? Per a què serveix? Des de quan es practica? Quines organitzacions el practiquen? Com es captura i emmagatzema el web? En la segona part es proposa una reflexió sobre l'aplicació de l'arxivament web des de la disciplina arxivística

Paraules clau: Preservació digital, arxivament web, arxivística, Internet, Biblioteques Nacionals, documents electrònics, tecnologies de la informació i la comunicació.

Title: Web archiving and the archival science

Abstract

Web archiving is a discipline that has its roots in Library Science and Information Science, and it is completely alien to the world of the archives in our country. The first part of this essay offers a brief overview on web archiving and, from an archival science point of view, will try to answer questions like: What is web archiving? What is it used for? When did it begin? Which organizations are putting it into practice? How can we capture and store a website? In the second part, we will try to think about the implementation of web archiving from the perspective of the archival science.

Keywords: Digital preservation, web archiving, records management, Internet, National Libraries, electronics records, information and communication technologies.

SUMARIO

Introducción

La voluntad de este trabajo es la de llamar la atención de la comunidad archivística hacia una de las ramificaciones de la preservación digital que más expansión se prevé que debe de tener en los próximos años, el archivo de la web, y que hasta ahora es una práctica inédita en nuestros archivos. Internet es un instrumento que desde hace años está experimentando un crecimiento que no conoce freno, y no sólo es el medio por excelencia a través del cual las personas intercambian todo tipo de bienes, de información, de datos, sino que también es un sistema de producción de documentos de primera magnitud.

Estos documentos que viajan por el ciberespacio, de máquina a máquina, de servidor a servidor, de página web a página web, no acostumbran a ser tratados archivísticamente. Con la consecuente pérdida de contextualización que eso comporta, y de garantías de fiabilidad, integridad y autenticidad que la disciplina archivística es capaz de garantizar. No se puede decir, en junio de 2013, que la gestión de los documentos electrónicos sea algo nuevo que se acabe de inventar, y a pesar de eso es una práctica que no está plenamente implantada en los archivos de Catalunya.

El tratamiento de las páginas web, su captura, gestión y preservación, es una faceta más de la gestión de documentos electrónicos que ha sido olvidada durante demasiado tiempo y puede que haya llegado el momento de empezar a pensar la manera en cómo los archiveros y gestores documentales podemos afrontarla. Pensemos, por ejemplo, en las sedes electrónicas que la Administración desarrolla para relacionarse con los ciudadanos. Estas páginas web contienen documentos con valor jurídico, generadores de derechos y obligaciones, y quizás sería positivo aplicar el archivo web a las sedes electrónicas para garantizar una mejor gestión de la información que contienen.

El presente Trabajo se dividí en dos partes, en la primera se intenta dar respuesta de forma breve y huyendo de tecnicismos, a cuestiones como en qué consiste el archivo de la web, cuándo se comienza a aplicar, de qué manera de lleva a cabo, o cuáles son las principales instituciones que lo practican hoy día. Intentamos entender también la peculiaridad de las páginas web, y cómo su naturaleza dinámica y su carácter efímero dificulta su tratamiento desde un punto de vista archivístico. Estos retos requieren nuevas respuestas tecnológicas que tienen que der desarrolladas bajo los criterios de la gestión documental para poder garantizar un correcto tratamiento de las unidades documentales que se crean, se intercambian o se almacenan en las páginas web.

La primer parte del trabajo finaliza con un breve repaso a los principales proyectos de archivo web del mundo, realizados por instituciones líderes que marcan la pauta del

desarrollo tecnológico y teórico, pero que no lo hacen bajo criterios archivísticos ya que en su mayor parte son bibliotecas nacionales. Comentamos también, y de forma destacada, el caso del PADICAT (Patrimoni Digital de Catalunya), de la biblioteca Nacional de Catalunya, un proyecto de referencia internacional desarrollado en el país.

La segunda parte del trabajo abandona el carácter expositivo e intenta relacionar la práctica del archivo web, creada y desarrollada desde la biblioteconomía y las ciencias de la información, con la disciplina archivística. Es, por lo tanto, un apartado más arriesgado en el que el autor intenta acercarse al *web archiving* desde una perspectiva archivística, destacando la importancia que tendría que la gestión documental empezara a tener en consideración las páginas web, sobre las que los archiveros deberían extender sus responsabilidades dado que será el principal medio de intercambio de documentos en un futuro cercano.

Intentamos dar una explicación, seguramente incompleta dada la complejidad de la cuestión pero meditada y sincera, de por qué el archivo de las páginas web es una práctica que no se realiza en los archivos de Catalunya, e imaginamos qué posibles escenarios de podrían dar en el futuro.

El presenta trabajo es, por lo tanto, una aportación humilde y modesta que quiere recordar a la comunidad archivística catalana la importancia de la preservación digital, y en concreto del archivo web, y la necesidad cada vez más acuciante de dar respuesta a los retos que la gestión de documentos electrónicos plantea. Cuestión hoy en día irrenunciable si no queremos perder al caballo de la innovación tecnológica, y que la archivística y gestión documental se conviertan en disciplinas olvidadas y arrinconadas debido a su incapacidad de aportar soluciones eficaces a los problemas que el nuevo entorno tecnológico genera.

PARTE I *WEB ARCHIVING*

1 ¿Qué es el *web archiving*?

El *web archiving* o archivo web es un conjunto de técnicas que tienen como finalidad la captura de páginas web. Es una práctica que se enmarca dentro de la preservación digital, que es la disciplina que se preocupa por la conservación a largo plazo de los objetos digitales, independientemente de su formato, del software con el que fueron creados, o del soporte físico que los contiene. La constante e imparable evolución de la informática ha generado un enorme abanico de sistemas para crear y procesar la información, y gran parte de este software y hardware ha quedado obsoleto con el paso del tiempo y ha sido substituido por otro. La preservación digital trabaja para evitar la pérdida de esta información y desarrolla diferentes técnicas (emulación, migración, captura de metadatos, etc.) para que estos objetos digitales sean accesibles para las sociedades del futuro.

Es importante darse cuenta de la velocidad con la que los objetos digitales pueden quedar desfasados, para hacernos una idea de una forma rápida basta con dar un vistazo al listado de formatos que recoge PRONOM[1], la base de datos de formatos creada por el departamento de preservación digital de los Archivos Nacionales del Reino Unido, que recoge más de 900, En la Wikipedia, en cambio, han compilado más de 3.500[2] formatos diferentes y es un listado incompleto. La principal característica de los documentos digitales es que se precisa de una máquina para poder leerlos[3], y si no disponemos de máquinas capaces de interpretar la codificación con la que los documentos fueron creados, esta información se ha perdido para siempre.

Para entender las dificultades de conservación de las páginas web tenemos que tener en cuenta que estas son una estructura de múltiples documentos, de diferentes formatos y tipos, interrelacionados entre ellos (bases de datos, formularios, documentos de texto, etc.) susceptible de sufrir problemas de obsolescencia si no reciben el debido tratamiento. Otros factores que determinan la vida de las páginas web son los relacionados con el servidor que les da alojamiento. Este servidor puede sufrir problemas técnicos, o extinguirse el contrato del servicio de mantenimiento de una web y causar la desaparición de una parte o la totalidad de la página.

Por lo tanto, la preservación digital es uno de los retos más importantes que debe afrontar la archivística moderna, que no sólo debe garantizar que los documentos electrónicos sean

[1] http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=listReport
[2] Colaboradores de la Wikipedia. *List of file formats (alphabetical)*. Wikipedia, The Free Encyclopedia, 2013 [data de consulta: 7 de març del 2013]. Disponible en <http://en.wikipedia.org/w/index.php?title=List_of_file_formats_(alphabetical)&oldid=537895641>
[3] ICA/ATOM pàg. 24.

auténticos, fiables, íntegros y completos, sino también accesibles. Hasta hora para garantizar la accesibilidad de los documentos al largo de la historia era suficiente con que el archivero los guardara en un lugar bien vigilado y tuviera cuidado de que no los deterioraran los insectos o los roedores. Hoy día, guardar los documentos electrónicos en un repositorio de acceso controlado no asegura su conservación por los problemas de obsolescencia que hemos comentado. Y es necesario llevar a cabo una seria de acciones preventivas que la preservación digital se encarga de estudiar.

El archivo de la web es, por lo tanto, la parte de la preservación digital que se encarga del estudio y desarrollo de técnicas, software t hardware para la captura de las páginas web y la conservación de estas a largo plazo.

Como veremos en el siguiente apartado, las primeras en comenzar a capturar páginas web fueron las bibliotecas nacionales y, por tanto, el término *web archiving* proviene del mundo de la biblioteconomía y las ciencias de la información, no de la archivística. Es importante que esta cuestión etimológica quede clara desde un principio.

El archivo de la web no sigue la metodología archivística ni tiene como objeto de estudio ni de trabajo los documentos de archivo, tal y como son entendidos en nuestra disciplina. Muy pocos archivos del mundo practican el archivo web, y se da por tanto la paradoja de que no son los archiveros los que archivan las páginas web sino los bibliotecarios y profesionales de la información. Puede parecer una cuestión menor, pero creemos importante resaltar este hecho para evitar confusiones, así como plantear desde aquí que, en aras de un lenguaje más claro, quizás sería mejor encontrar otra expresión que evite la ambigüedad de la expresión archivo web. Y más aún si tenemos en cuenta que la archivística por un lado y la biblioteconomía por otro son disciplinas diferentes pero con algunas fronteras compartidas, que han hecho esfuerzos por delimitar su campo de estudio y dejar claras sus competencias.

2 ¿Cuándo se comienza a archivar la web?

Las páginas web se comienzan a capturar cuando se populariza su uso y creación de la mano de la expansión de Internet. Oficialmente se considera que el nacimiento de Internet se produce el 1 de enero de 1983[4], con la primera red de largo alcance basada en tecnología TCP/IP que puso en funcionamiento la *National Science Foundation* de los Estados Unidos. Pero no es hasta la década de los noventa que se populariza su uso a partir de dos hitos imprescindibles. Por un lado, el nacimiento en 1991 del proyecto *World Wide Web*. I por el otro, el trabajo del ingeniero británico Timothy Berners Lee, que en 1993 desarrolló el lenguaje de marcas HTML e impulsó la creación del protocolo de intercambio http.

A partir de ese momento empiezan a proliferar las páginas web que cualquier ciudadano del mundo puede generar, a medida que va aumentando el acceso a Internet y, consecuentemente el contenido de la red, se despierta la preocupación sobre cómo gestionar tremenda cantidad de conocimiento. Las bibliotecas nacionales, que tradicionalmente habían atesorado la totalidad de publicaciones del propio país mediante la gestión del depósito legal, ponen en marcha iniciativas para continuar desarrollando esta misión propia en el mundo digital.

De esta manera durante el año 1996 se configuran tres proyectos pioneros, *Kulturaw3* de la Biblioteca Nacional de Suecia, *Pandora* de la Biblioteca Nacional de Australia, y el *Internet Archive*, que se funda en San Francisco (California) pero que no es una iniciativa nacional sino una entidad sin ánimo de lucro con vocación internacional.

Posteriormente, las bibliotecas nacionales de otros países fueron creando sus propios proyectos de archivo web y los repositorios digitales donde conservar la información resultante. Como, por ejemplo, el archivo web de Nueva Zelanda, el archivo web de la Biblioteca del Congreso de los Estados Unidos, el archivo web de Noruega o, ya en nuestra casa, el PADICAT, el Patrimoni Digital de Catalunya, creado por la Biblioteca de Catalunya el 1995.

Por tanto, la historia del *web archiving* es muy corta, unos diecisiete años des de la puesta en marcha de los proyectos pioneros, pero muy activa dado que en un período de tiempo tan breve han nacido más de 40 proyectos de archivo web en 26 países diferentes[5] que han capturado más 7.500 terabits de información.

[4] http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet
[5] Gomes, Daniel

3 ¿Por qué archivar la web?

Obligación legal

Como hemos visto anteriormente, el impulso inicial de archivar las páginas web proviene de las bibliotecas nacionales, que tienen como uno de los pilaras básicos de su misión institucional recopilar la producción bibliográfica, discográfica, etc., del país, conservarla y hacerla accesible a los ciudadanos. El listado de obras de obligatorio registro en el Depósito Legal es muy variable[6] y va desde partituras musicales hasta los carteles publicitarios que podemos ver en las carreteras. Las obras digitales que se producen en Internet también son de interés para el Depósito Legal, y es obligación de las bibliotecas nacionales hacerse cargo de ellas. El archivo del web ofrece a las bibliotecas la posibilidad de capturar este patrimonio del país ya sea para dar cumplimiento a la obligación legal en sí, o para desarrolla la vocación de defensa y promoción de la cultura que tradicionalmente ha caracterizado a estas instituciones.

Investigación

Con el desarrollo de Internet se ha producido una rápida transformación de la manera en como nuestra sociedad se relaciona y comunica. La expansión del acceso a la red ha posibilitado que Internet se convierta en el medio principal en el que se da todo tipo de intercambio, desde los puramente económicos (transacciones bancarias, compra y venta de productos, etc.) hasta los puramente informativos, laborales, sociales, etc. Internet ha transformado la manera en la que gran parte de la humanidad de comunica, i es un reflejo directo de la actividad de las comunidades que a través de ella participan. Este aspecto se ha visto reforzado por el estallido de las redes sociales, verdaderas ventanas donde los individuos exponen y retransmiten su vida en tiempo real.

Todo este enorme volumen de información es la materia prima con la que los historiadores de la sociedad del siglo XXI tendrán que trabajar. Y para que eso pase será imprescindible la conservación de estas páginas web que testimonian dicha vitalidad. Pero también los investigadores de las más diversas disciplinas necesitarán el archivo web para su posterior consulta y estudio: sociólogos, antropólogos, economistas, especialistas en márquetin, periodistas, historiadores del arte y un largo etcétera tendrán en un futuro –y tienen ya hoy día- en Internet su campo de estudio.

[6] http://www.bnc.cat/Professionals/Diposit-legal#quinesobres

Memoria universal del conocimiento

Por otro lado, también es necesario destacar que existe una línea de pensamiento estrechamente ligada al nacimiento y desarrollo de Internet, que defiende una serie de principios como el de la neutralidad en la red, los derechos digitales (derecho de acceso a Internet, libertad de expresión y de información en la red, etc.), que considera Internet un bien universal. Es una filosofía muy cercana a los principios de la Wikipedia, por ejemplo, en la que grandes comunidades de individuos trabajan de forma desinteresada para el bien común de la compilación del conocimiento.

Desde este punto de vista, Internet es una fuente de información y conocimiento universal a la que todo el mundo debería poder acceder, incluidas las generaciones futuras. Y el archivo de la web se presenta como una de las grandes estrategias de preservación digital que puede garantizar que este patrimonio de la humanidad no se pierda. Con este espíritu nació el archivo digital más grande hoy día el *Internet Archive*, que da acceso libre a todo su depósito.

Estos tres conceptos son argumentos poderosos que justifican el ejercicio del *web archiving* pero no tienen por qué ser los únicos. Otro motivo más mundano, si se quiere llamar así, es el interés que cada productor de una página web puede tener en conservar la información publicada en línea para su propio uso, ya sean empresas, universidades, instituciones, administraciones públicas o particulares.

4. ¿Cómo se archiva la web?

La finalidad del presente trabajo no es tratar en profundidad el aspecto técnico del *web archiving*, pero sí querríamos apuntar brevemente la manera en la que se realiza para proporcionar una idea general de su metodología a nivel intelectual sobretodo, para el lector que se acerca por primera vez a esta disciplina.

4.1. Crawlers

La captura de las páginas webs se realiza mediante un software conocido con el nombre de *web crawler*. Es un programa automatizado (*internet bot*) que tiene como finalidad buscar, capturar e indexar las url de las páginas web. Por lo tanto, el contenido que se captura de Internet depende en gran medida de las instrucciones que se le dan a los *crawlers*.

De la gran cantidad de *crawlers* existentes podríamos destacar *Heritrix*, un programa de código abierto desarrollado por *Internet Archive* y las bibliotecas Nacionales Nórdicas. Algunas de las organizaciones que lo utilizan son instituciones que lideran el panorama del archivo web como la Biblioteca Nacional de Australia, la Biblioteca Nacional de Francia o la Biblioteca Británica.

Otro *crawler* que podríamos destacar es *HTTrack*. Este fue desarrollado por Xavier Roche y permite descargar las páginas capturadas en el propio ordenador. Otro capturador web que podríamos mencionar es *Wget*, muy ligado al sistema operativo GNU[7].

Cabe destacar también que la proliferación de empresas de servicios de captura web ha dado a luz toda una amplia gama de capturadores que no entraremos a comentar aquí para no salirnos del tema principal. Solamente querríamos apuntar que tres de los más conocidos son *Archive-It*[8], *Archivethe.net*[9] i *Hanzo archives*[10].

4.2. Modelos de captura

En función de los criterios que se utilicen para llevar a cabo la captura y el archivo de las páginas web podemos establecer tres enfoques diferentes:

[7] http://www.gnu.org/gnu/gnu.html
[8] http://www.archive-it.org/
[9] http://archivethe.net/fr/
[10] http://www.hanzoarchives.com/

-Modelo integral o exhaustivo: pretende la captura automatizada de todas las páginas web bajo un criterio general, ya sea lingüístico (por ejemplo, todas las páginas en catalán), de ubicación de los servidores que alojan la web (por ejemplo, páginas alojadas en servidores ubicados en Catalunya), de dominio (por ejemplo, todas las páginas con el dominio ".cat"), etc.

Este modelo asegura la compilación de gran cantidad de información pero no puede acceder a la Internet invisible, que es esa parte de Internet que se genera de forma dinámica cuando un usuario hace una consulta, o que está protegida con contraseña (en el punto 5.1. lo tratamos con mayor profundidad). Esta manera masiva de capturar páginas web no permite un tratamiento detallado de la información recopilada y, por lo tanto, presenta más limitaciones en el momento de ser posteriormente recuperada.

-Modelo selectivo: sigue unos criterios que acostumbran a ser temáticos y, antes de efectuar la captura, la entidad llega a un acuerdo con los editores y productores de las páginas web. Es, por tanto, una tarea más lenta que requiere contactar con productores de recursos web, negociar las condiciones y establecer acuerdos pero, en cambio, la información capturada se puede tratar al detalle y asegurar una explotación posterior más eficiente. Es un sistema costoso teniendo en cuenta los recursos humanos necesarios para llevarlo a cabo.

-Modelo híbrido: consiste en una combinación de los dos modelos anteriores. Muchos proyectos nacionales, como por ejemplo el PADICAT, lo aplican e intentan capturar la web de forma sistemática pero también llegando a acuerdos con las instituciones productoras. Puntualmente se hacen capturas selectivas enfocadas a la celebración de algún acto social o histórico destacado, como por ejemplo al celebración de unos juegos olímpicos o un proceso electoral.

4.3. Políticas

El archivo web no consiste en poner a un robot a capturar páginas web, sino que antes de esto, y también con posterioridad a la captura, hay toda una serie de decisiones que las organizaciones tienen que tomar. Son unas decisiones con las que los archiveros estamos familiarizados ya que son muy similares a las normas de gestión documental que recomienda la ISO-15489-1 y la ISO-15489-2. De una manera muy esquemática podemos dividirlas en los siguientes apartados:

Políticas de captura

Cada institución establece los parámetros de sus capturadores en función de la misión y de los objetivos propios. Debe tener en cuenta la ley de propiedad intelectual y la de contenidos digitales de su país, que le indicará los límites de captura que no deben ser excedidos. También ha de tener en cuenta el presupuesto de la institución así como las soluciones técnicas que utilizará para la captura.

Política de gestión y acceso de las colecciones

Son todas aquellas decisiones relativas al tratamiento de las páginas web capturadas: ordenación, clasificación, indexación, etc. Será necesario establecer quiénes son los responsables del tratamiento y determinar si el contenido será de acceso abierto o bien tendrá alguna restricción y, en este caso, identificar quiénes son los sujetos que podrán acceder.

Política de preservación

Mediante la cual debe determinarse el contenido que hay que conservar de manera permanente y el que debe ser eliminado. En consecuencia debe fijarse un calendario de eliminación por una parte, y por otra las técnicas de preservación que se aplicarán (migración, emulación, etc.) en función de las características técnicas de las colecciones. Otro aspecto a tener en cuenta es el de la seguridad y mantenimiento del repositorio digital en el que las colecciones se conservan, y los requisitos que debe cumplir con tal de garantizar su confiabilidad.

5 Obstáculos del *Web Archiving*

Como consecuencia de la propia naturaleza de Internet el archivo web se encuentra con toda una serie de problemas y limitaciones de muy diversa índole que podríamos comentar brevemente en cuatro apartados.

5.1. Dificultades técnicas

Crecimiento exorbitante

En primer lugar hay que destacar que Internet es un espacio que no para de crecer, es finito pero está en constante transformación y cambio, y crece a un ritmo superior a la velocidad con la que se captura la web. Con sólo plantearse dos preguntas se comprende fácilmente esta diferencia de ritmo: ¿Cuántos agentes hacen crecer internet? ¿Cuántos agentes capturan la web? Los segundos nunca podrán igualar en número a los primeros y la idea de si se podría llegar a hacer un *backup*, una copia de seguridad, de todo Internet queda fácilmente despejada. Otra pregunta es si es necesario replicar todo Internet, para los archiveros, que evalúan la documentación que generan las instituciones en las que trabajan y conservan solamente una parte de ella, no es necesario. Pero algunas de las organizaciones que actualmente archivan el web sí tienden a capturar una gran parte de Internet, como hemos visto, y tienen que hacer frente al reto de esta gran expansión.

Naturaleza dinámica y carácter efímero de la web

Por otro lado, las páginas web, como objetos digitales, son bastante especiales ya que en su mayoría no constituyen un único documento electrónico sino que acostumbran a ser una arquitectura de múltiples documentos, de diferentes formatos y tipos, interrelacionados entre ellos. Además tenemos que tener en cuenta el carácter dinámico de la web, que es la característica que lo convierte finalmente en un objeto tremendamente complejo.

Brügger[11] indica que el dinamismo de la web se da de tres maneras diferentes:

-Desde el punto de vista de quien envía la información:

a) Dinámica de actualización. Una web creada puede ser modificada en cualquier momento.

b) Dinámica de proliferación. En cualquier momento se pueden crear nuevas webs.

-Desde el punto de vista de los elementos textuales (ya sea texto, imágenes, sonidos, etc.):

[11] Brügger, Niels. Capítol 3.

a) Dinámica del movimiento. Entre elementos textuales y dentro de los mismos, ya sea con enlaces que redireccionan al visitante, o descargas que se ejecutan cuando se clica sobre un botón preconfigurado.

b) Dinámica de la complejidad. A medida que una web integra más y más elementos interrelacionados entre ellos y con las acciones que realizará el visitante, aumenta su complejidad.

-Desde el punto de vista del recipiente que contiene la información

a) Dinámica del equipamiento. La web se visualizará de forma diferente en función del navegador que la ejecute o del dispositivo (teléfono móvil, tableta táctil, ordenador personal, etc.)

b) Dinámica de las acciones. Las acciones que un visitante hace sobre la web la pueden personalizar, per ejemplo mediante las *cookies*, y esto hace que cambie su comportamiento en función del visitante.

Un aspecto ligado al dinamismo de la web es su carácter efímero, las páginas web en constante actualización eliminan parte de sus contenidos con elevada frecuencia. Pensemos en una web de noticias: ¿cuántas veces al día cambia la portada de un periódico digital? Cuando se captura una web se pretende reproducir una pequeña parte de la realidad que existe en Internet en un repositorio digital, pero el contenido guardado dejará de tener una equivalencia con la realidad cuando la página se modifique, y solamente quedará al captura como testigo de un momento concreto. Incluso se puede dar el caso de que la captura registre una realidad que nunca tuvo lugar. Es lo que pasaría si una página se actualiza mientras se está capturando. Siguiendo con el ejemplo del periódico digital, la portada capturada mezclaría noticias de antes y de después de la actualización, y sería un híbrido que no reflejaría la realidad con fidelidad.

Por lo tanto, las páginas web están constantemente cambiando o desapareciendo, y esto significa que una enorme cantidad de información se está perdiendo para las generaciones futuras.

Internet profunda o *deep web*

Este término hace referencia a la parte de Internet que no pueden indexar los motores de búsqueda habituales y que, por lo tanto, está fuera del acceso de la mayoría de internautas.

Tampoco pueden acceder a la Internet profunda los *crawlers* que se utilizan para la captura web.

Los principales motivos por los cuales los motores de búsqueda no pueden acceder a la Internet profunda son tres:

-Algunos editores web protegen sus páginas para evitar que sean indexadas, ya sea con contraseñas, cortafuegos o ficheros de robots .txt.

-Gran parte de los recursos de Internet requieren de una consulta a la base de datos para poder acceder a su información. Mientras esta no se haga el contenido no se hace visible (por ejemplo, las webs de diccionarios o enciclopedias).

-El contenido se encuentra en un formato no indexable.

Aguillo[12] establece que la Interne profunda está formada por cuatro grupos de recursos digitales:

-Bases de datos bibliográficas: formadas por catálogos de biblioteca accesibles a través de una pasarela web, otras bases de referencias bibliográficas y catálogos de bibliotecas.

-Bases de datos alfanuméricas: todas las bases de datos que no tengan carácter bibliográfico y otros recursos que requieren de una pasarela web para su consulta (enciclopedias, diccionarios, etc.).

-Archivos y revistas electrónicas: acostumbran a ser recursos de gran calidad a los cuales se accede exclusivamente con una consulta previa.

-Ficheros no HTML o textuales: como por ejemplo el Acrobat PDF.

Estándars web y formatos de software cambiante

La vertiginosa evolución de la tecnología produce nuevos formatos, nuevos programas y nuevas maneras de concebir, diseñar e implementar páginas web. Este factor dificulta la tarea del archivo web, que debe ir adaptándose constantemente a estos cambios y debe hacer evolucionar sus propios sistemas de captura y almacenamiento para adaptarse a las nuevas situaciones.

[12] Aguillo, Isidro F.

También la evolución de los lenguajes de marcas y de los estándares web requiere una adaptación de los agentes que capturan la web. Por motivos como estos es importante la tarea del *International Internet Preservation Consortium*, que promueve el desarrollo y uso de herramientas comunes, técnicas y estándares que faciliten la creación de archivos web.

La esperanza de la Web semántica

El proyecto de construir una web semántica que facilite la interoperabilidad de los sistemas informáticos y facilite la interpretación de la información por parte de las máquinas quizás sea muy útil para el archivo web. Este proyecto, ideado per Tim Berners-Lee, promueve el uso de unos lenguajes de marcas estándares para describir la información *on line* de tal manera que las páginas web sean ,más comprensibles para las computadores y estas puedan operar con el significado de la información de forma automática.

Sin duda sería un gran avance a la hora de consultar los archivos web, ya que la conexión automática y significativa entre los datos proporcionaría al usuario resultados mucho más ricos que la mayoría de buscadores de hoy día.

5.2. Dificultades legales

Otro obstáculo que deben gestionar las entidades que practican el archivo web es la complicada cuestión de los derechos de autor de los contenidos de Internet y el conflicto que pueda derivarse de la protección de datos personales. Ante esta cuestión no es posible una estrategia estandarizada debido a que en cada país existe una reglamentación diferente sobre estas materias y es necesario realizar un enfoque diferente para cada situación.

Son tres las posibles estrategias que se suelen aplicar al respecto:

-Capturar y archivar exclusivamente webs libres de derechos de autor.

-Desarrollar políticas efectivas de gestión de derechos.

-Capturar masivamente pero permitir el acceso sólo en función del solicitante. Es lo que se hace, por ejemplo, en Dinamarca, donde la ley de depósito legal permite desde 2005 capturar la parte del web deseada sin necesidad de pedir permiso a los propietarios. Por el contrario, para no vulnerar los derechos, no se ofrece acceso a estas capturas al público general.

5.3. Dificultades de organización

En este punto querríamos comentar el hecho de que Internet no está gobernada por ninguna entidad y crece y se desarrolla de manera descentralizada. Aunque son muchos los agentes y lobbies que intentan extender su control mediante forzando la aprobación de leyes y reglamentos, en realidad no se puede decir que Internet sea un espacio que pertenezca a una única entidad. Este hecho positivo, que Internet no tenga un dueño, tiene su contrapunto negativo des del punto de vista del archivo de la web, y es que los estándares y las recomendaciones que facilitarían la captura no se pueden imponer y son responsabilidad de cada individuo y de cada institución productora.

Por otro lado, existen diferentes maneras de afrontar el archivo web dependiendo de la institución que lo lleve a la práctica, ya sean Bibliotecas Nacionales, museos, centros de historia, galerías de arte, etc. Cada uno tendrá su propia estrategia y sus propios objetivos, pero para avanzar en el campo de la preservación digital del contenido *on line* sería necesario reforzar la cooperación entre los diferentes agentes, para ahorrar esfuerzos y no repetir capturas.

5.4. Dificultades económicas

El gesto técnico de capturar y archivar una página web no tiene por qué ser una operación muy costosa económicamente hablando. Existen programas libres para la captura, y es posible utilizar el disco duro del propio ordenador como unidad de almacenamiento, o incluso el espacio que muchas empresas ofrecen gratuitamente en la nube.

Pero, como hemos visto, el sentido último del archivo web tal y como lo aplican las bibliotecas es el de salvaguardar grandes cantidades de información para evitar que se pierdan, y esta es una tarea imposible de llevar a cabo de manera individual y requiere del liderazgo de grandes proyectos de captura y almacenamiento. Y es en este sentido en el que la situación cambia radicalmente y los costes empiezan a dispararse. A parte de un software de captura es necesario un programa de gestión que posibilite trabajar con la información, indexarla, aplicar políticas de acceso, de preservación, etc. Sería necesario disponer de un sistema para hacer accesible las capturas, con una interface que permitiera la búsqueda y la consulta. Sería imprescindible un equipo de personas expertas en la materia que manipularan los instrumentos de captura, edición y tratamiento de la información recopilada.

Sería ineludible disponer de un equipo informático especialmente diseñado para almacenar grandes cantidades de información que cuenten con las medidas de seguridad suficientes

para evitar ataques externos, pero que al mismo tiempo ofrezca acceso en línea y permita la consulta a distancia. Deberían implementarse sistemas de copias de seguridad periódicas para garantizar la perdurabilidad de la información, etc. Como se puede imaginar los costes económicos derivados de la puesta en marcha y el mantenimiento de toda estructura tecnológica y humana son muy elevados, Miquel Térmens[13] concluye que la evolución tecnológica comportará un abaratamiento de los costes derivados del hardware y el almacenamiento digital, pero que el verdadero reto de estos grandes repositorios digitales masivos no es su mantenimiento en el tiempo sino el hecho de poder disponer del capital suficiente como para iniciar el proyecto

Por lo tanto, es clave una correcta planificación previa, y totalmente recomendable, casi se podría decir que imprescindible, buscar alianzas entre organismos e instituciones para poder llevar a cabo de forma coordinada y compartida proyectos de captura web sólidos, robustos y duraderos.

[13] Térmens, Miquel

6 ¿Quién archiva la web?

Como ya hemos visto quienes comenzaron a desarrollar el archivo web fueron las bibliotecas nacionales, pero durante los diecisiete años de historia esta técnica los cambios tecnológicos y organizativos han hecho que actualmente encontremos más actores con un papel destacado dentro de este escenario.

Actualmente existen más de sesenta proyectos de archivo web en todo el mundo, sería muy extenso hablar de todos ellos y estaría fuera de las intenciones de este trabajo. Por lo tanto lo que hemos hecho es establecer cuatro categorías en las que creemos que se clasificar estos actores, y comentar brevemente algunos de sus ejemplos más representativos.

6.1. Bibliotecas nacionales

Son las instituciones que primero encararon la problemática de capturar las páginas web y las que primero desarrollaron soluciones tecnológicas para poder llevarlo a cabo. El hecho de ser instituciones únicas y relevantes con un peso específico en su país, ha posibilitado que pudieran contar con presupuestos suficientes como para arrancar sus respectivos proyectos. En muchos casos son la única institución del país que trata este tema y eso, de alguna manera, las ha convertido en proyectos nacionales y en actores de referencia.

Los proyectos de las bibliotecas nacionales tienen como denominador común la búsqueda de aquella parte de Internet producida por el propio país o que de alguna manera le pertenezca atendiendo a criterios históricos, sociales o culturales. Aun así, no es una tarea sencilla determinar qué parte de Internet pertenece a un determinado país cuando la red es un vasto espacio desprovisto de fronteras.

Una de las pioneras es la **Biblioteca Nacional de Suecia**, el archivo web de la cual recibe el nombre de ***Kulturaw3***. No es extraño que esta institución fuera de las primeras en preocuparse por esta cuestión, y es que Suecia ha sido muy precoz en esta materia y ya en 1661 encargó a la Biblioteca Real las tareas de coleccionar todas las publicaciones impresas en el país. Por lo tanto, les viene de lejos esta preocupación de atesorar la propia producción cultural.

En 1996 inauguran el proyecto *Kulturaw3* y comienzan a plantearse cuestiones como la de definir que parte de Internet se podría considerar que pertenece a Suecia. Mediante una estrategia exhaustiva capturan todos los servidores con dominio sueco (.se), pero también los servidores con cualquier otro dominio registrados en Suecia (ya sea con una dirección sueca, o con un número de teléfono sueco). También realizan la captura de periódicos *on line*, a diario más de 140. A pesar de todo, la legislación del país no permite el acceso al

contenido de *Kulturaw3* por Internet, y sólo se puede consultar la colección desde la misma Biblioteca Nacional de Suecia.

Otra gran iniciativa pionera fue la de **PANDORA**, el archivo de la **Biblioteca Nacional de Australia** que se inició el año 1996. Hay que decir que no es el único proyecto del continente de estas características (se puede mencionar también *Our Digital Island* de Tasmania, y *Territory Stories* de la Biblioteca del Estado de *Northern Territory*) pero sí es el más importante y uno de los que marcó la pauta del desarrollo del archivo web tal y como lo conocemos hoy día.

PANDORA (*Preserving and Accessing Networked Documentary Resources of Australia*) ha construido una red colaborativa formada por nueve piezas, las bibliotecas de los seis estados continentales más el Memorial Australiano de la Guerra, el Archivo Nacional de Imagen y Sonido, y el Instituto Australiano de Estudios Aborígenes y de los Habitantes del Estrecho de Torres.

PANDORA es un ejemplo de archivo web selectivo que no pretende capturar toda la web nacional sino que hace una selección entre aquellos recursos que considera más valiosos y que deben ser preservados. Cada uno de los socios miembros tiene sus propias prioridades de captura, pero todos trabajan desde esta estrategia selectiva y archivan materiales relativos a la vida social, cultural, política e intelectual de los australianos. Esto incluye blogs, páginas web estatales, de organizaciones, colecciones de periódicos y capturas destinadas a registrar el impacto de un determinado acontecimiento.

Han desarrollado su propio programa de tratamiento, PANDAS, que les permite gestionar las colecciones de forma detallada, indexando, catalogando y controlando el acceso a los contenidos. Esta manera de trabajar pormenorizada es bastante lenta y no les permite capturar todos los recursos que querrían preservar. Para evitar que estos se pierdan contratan los servicios del *Internet Archive* que es capaz, con su programa *Heritrix*, de capturar un enorme espacio web en muy poco tiempo, eso sí, son el detalle que PANDAS proporciona.

El acceso al contenido de PANDORA es totalmente libre para cualquier persona con conexión a la red, excepto una pequeña parte (aproximadamente el 2%) que casi en su totalidad se puede consultar desde los ordenadores de la misma biblioteca.

El último proyecto nacional del que hablaremos es el **Archivo Web del Reino Unido** que es un ejemplo de la capacidad de coordinación de múltiples instituciones, en su mayoría bibliotecas nacionales, para construir un gran repositorio digital representativo de la importancia histórica, social i cultural del país. El Archivo Web del Reino Unido fue

constituido en el 2004 por la **Biblioteca Británica** y a lo largo de su trayectoria ha colaborado con diversas instituciones. Algunos socios permanentes son la Biblioteca Nacional del País de Gales, JISC (*Joint Information System Committee*), *The Welcome Library* o la Biblioteca de les Dones de la *Metropolitan University* de Londres.

El Archivo Web del Reino Unido llevaba a cabo una estrategia selectiva de captura web, negociando con las organizaciones y empresas propietarias antes de capturar su web, pero durante el desarrollo del presente trabajo ha habido un cambio legislativo en materia de depósito legal en el Reino Unido que ha modificado el escenario. Des del 6 de abril de 2013 una nueva normativa amplia las capacidades para capturar y archivar la web a seis bibliotecas del país: la Biblioteca Británica, la Biblioteca Nacional de Escocia, la Biblioteca Nacional del País de Gales, la Biblioteca Bodleiana, la Biblioteca de la Universidad de Cambridge y el *Trinity College* de Dublín.

Desde la biblioteca Británica se ha lanzado una campaña en Internet y en las redes sociales pidiendo la participación de la sociedad británica para definir los criterios de captura i crear un debate en torno a qué páginas web deben preservarse y qué utilidades se les puede dar. Asistimos, por tanto, a una nueva etapa en la que el Archivo Web del Reino Unido verá incrementado exponencialmente su contenido y veremos de qué manera los instrumentos institucionales de comunicación conseguirán difundir en la ciudadanía el conocimiento del *web archiving* y de los beneficios que este puede tener para la sociedad y el mundo empresarial británico.

Existen muchos más proyectos de archivo web desarrollados por bibliotecas nacionales pero no tenemos tiempo de comentarlos todos, algunos son la Biblioteca Nacional de Francia, el Archivo web del Gobierno de Canadá, el Archivo web croata, el Archivo web islandés, el Archivo web de Eslovenia, el Archivo web de Austria, etc.

6.2. Proyectos supranacionales

Entendemos por proyectos supranacionales aquellos que no se circunscriben bajo los intereses de un país. Proyectos cuyos intereses podríamos considerar transnacionales, que buscan alcanzar una misión que supera las fronteras.

El primero de estos proyectos que queremos comentar tanto porque fue pionero en el campo del archivo web como porque es el primero del mundo en cantidad de información capturada, es el ***Internet Archive***. Es una fundación sin ánimo de lucro que fue creada en el año 1966 en San Francisco (EE.UU.) con la finalidad de construir la biblioteca de Internet, que recopilara todo el conocimiento y la cultura a través de la red para evitar su pérdida, y

proporcionar acceso a las generaciones futuras y a los investigadores, historiadores y estudiantes del presente.

El *Internet Archive* recopila webs públicas de todo el mundo, de todos los dominios, en cualquier idioma y proporciona acceso público a este contenido. Su capacidad de captura es muy elevada y posee el archivo web más grande del mundo que crece a un ritmo de 100 terabytes al mes. Colabora con un gran número de instituciones que le encargan la captura y archivo de sus propias webs. Y otras, especialmente universidades, utilizan sus herramientas informáticas y su repositorio para crear y gestionar sus propias colecciones.

Pero el *Internet Archive* no sólo archiva webs sino todo aquello que denomina artefacto cultural, y permite que cualquier persona en cualquier parte del mundo pueda subir contenido a su repositorio. Por lo tanto, allí podemos encontrar películas, programas de televisión, programas de radio, e-books, etc. El *Internet* Archive también trabaja para incentivar la creación de archivos web y, mediante la plataforma *Archive-it*, ofrece servicios de captura a cualquiera que tenga interés en utilizarlos. Y organiza cada año el *K12 Web Archiving* que es una formación destinada a que los jóvenes estudiantes aprendan a crear sus propias colecciones de páginas web.

Otra organización de referencia en materia de archivo web es ***International Internet Preservation Consortium*** (IIPC), una institución internacional que trabaja para la preservación del contenido de Internet. Fue fundado en julio de 2003, y ha pasado de contar con doce miembros iniciales a la treintena que aproximadamente hoy día lo constituyen.

EL IIPC da apoyo a las bibliotecas nacionales para que inicien proyectos de archivo web, trabaja para crear herramientas y técnicas comunes que posibiliten la creación de archivos de páginas web. Promueve los estándares entre sus organismos miembros para sumar esfuerzos y, de esta forma, hacer evolucionar los métodos de captura desarrollando recursos tecnológicos cada vez más especializados. También es una plataforma que coordina diferentes grupos de trabajo que estudian aspectos como los métodos de captura, las políticas de acceso al contenido y los retos de la preservación digital. El IIPC publica periódicamente los resultados de su trabajo y organiza conferencias y reuniones de expertos internacionales. También gestiona una lista de correo electrónico donde se debaten todas estas cuestiones.

Por último comentaremos el caso de la ***Internet Memory Foundation*** que es una fundación sin ánimo de lucro con sede en París y Ámsterdam cuya misión es la preservación de Internet. Colabora con múltiples organizaciones y ofrece apoyo para la creación de colecciones digitales que llevan a cabo instituciones europeas como la Biblioteca Nacional

Central de Florencia, el Parlamento del Reino Unido o la Organización Europea para la Investigación Nuclear (CERN).

Internet Memory Foundation está involucrado en diversos proyectos de alcance europeo, muchos de ellos fundados por la Comisión Europea, que pretenden desarrollar tecnología y metodología para la próxima generación de herramientas de archivo web. En definitiva, es otro agente importante en el desarrollo y difusión del archivo web.

6.3. Universidades y organismos de investigación

Otras instituciones que trabajan el archivo web son las entidades vinculadas al campo de la investigación. El mundo universitario en Estados Unidos es muy competitivo y las universidades que disponen de un elevado presupuesto llevan a cabo proyectos de archivo web, que enriquecen su patrimonio digital y las convierten en instituciones de referencia a las que acuden investigadores atraídos por la relevancia de sus colecciones. De esta forma las universidades que lideran productos de archivo web no sólo están garantizando el acceso a un contenido que de otra manera se hubiera perdido para siempre, sino que también se están convirtiendo en sujetos activos que utilizan el material capturado para crear nuevo conocimiento en beneficio de la comunidad cuando ofrecen sus colecciones digitales para que los investigadores y expertos las estudien y exploten.

Es el caso de la **Biblioteca de la Universidad de Columbia** que utiliza la plataforma *Archive-it*, desarrollada por *Internet Archive*, para crear y gestionar sus propias colecciones. Las webs capturadas pertenecen a algunos de los departamentos de la Universidad o a sus afiliados, o bien son páginas web de individuos o instituciones que tienen documentos o trabajos custodiados en el archivo físico de la institución.

Otro ejemplo es el **Servicio de Colecciones de Archivos Web** de la Universidad de Harvard, que nació en julio de 2006 y fue el primer proyecto fundado por la *Library Digital Iniciative* que estaba orientado al estudio de la preservación a largo plazo de material nacido digital. Algunas de las colecciones que gestionan tratan temas como la vida y estudio en Harvard, la historia de las mujeres en Estados Unidos, o el debate creado en Japón entorno a la revisión de la Constitución vigente después de la Segunda Guerra Mundial.

Por otro lado, la Universidad de California mediante su **Biblioteca Digital de California** ofrece servicios de archivo web a una larga lista de colaboradores, en su mayoría relacionados con el mundo universitario y de la investigación, para que puedan crear, archivar y gestionar sus propias colecciones digitales. De esta forma los socios de la Biblioteca Digital de California, algunos tan destacados como las Bibliotecas de la

Universidad de Nueva York o la biblioteca de la Universidad de Berkeley, pueden capturar y mantener siempre accesibles sus propias páginas web. Y al mismo tiempo crear colecciones temáticas sobre aquellas materias en las que sean una referencia en el mundo de la educación y el conocimiento.

6.4. Empresas

Entre las empresas que practican el archivo web podemos establecer una clara división entre aquellas que se dedican a ofrecer servicios de archivo a terceros, y aquellas que usan la captura para archivar sus propias páginas web.

Desde el origen del archivo web en 1996 y con la evolución tecnológica que se ha ido produciendo han crecido una serie de empresas que hacen de la captura, el archivo y la gestión de páginas web su campo de negocio. Son empresas especializadas en sistemas de información y en la comunicación por Internet que ponen sus herramientas tecnológicas al servicio de sus clientes. Algunas de ellas son *Hanzo Archives*, *Nextpoint* o *Smarsh Web Archiving*.

Por otro lado, el porcentaje de empresas que realiza capturas de su propia web es mínimo pero seguro que tendrá un crecimiento importante durante los próximos años. Especialmente entre aquellas empresas que utilicen su página web para comunicar al mundo sus servicios, y entre aquellas que la utilicen como un instrumento de relación con sus clientes. Otras webs que será interesante capturar serán aquellas que sirvan para intercambiar información relevante (empresa-cliente), o aquellas que ofrezcan servicios de oficina al estilo sede electrónica.

7 El archivo de la web en Catalunya: el PADICAT

A mediados de la pasada década la Biblioteca de Catalunya hace un replanteamiento de sus mecanismos y estrategias más adecuadas para preservar y difundir el patrimonio nacional en un mundo que ofrece nuevas posibilidades tecnológicas. Fruto de esta reflexión el año 2005 emprende cuatro proyectos innovadores:

-Colaboración con *Google Books* para la digitalización de una parte del fondo bibliográfico.

-*Memòria Digital de Catalunya*, es un repositorio cooperativo y de acceso abierto que contiene colecciones digitalizadas relacionadas con Cataluña y su patrimonio.

-*Arxiu de Revistes Catalanes Antigues (ARCA)*: digitalización de revistes catalanes que ya no se publican.

-*Patrimoni Digital de Catalunya* (PADICAT), el archivo web de Cataluña.

En 2005 arranca la fase de nacimiento del PADICAT, en la que se dan los pasos preliminares de estudio y preparación del proyecto. En julio del 2006 se comienza a capturar la web de manera automatizada, y el momento culminante llega en septiembre del mismo año con el estreno del portal web en el que se da acceso abierto al material compilado.

Durante la fase de crecimiento (2007-2008) se trabaja para ejecutar los programas de desarrollo trazados con anterioridad y conseguir un funcionamiento normal de todas las piezas del engranaje: captura, indexación, archivo, puesta a disposición pública. En la fase de consolidación (2009-2011) se lleva a cabo un destacable trabajo de difusión entre las instituciones del país que da como resultado la firma de convenios de colaboración con más de 450 organismos que cooperaran para que el PADICAT capture sus páginas web. EL 11 de septiembre de 2011 se moderniza el portal web de acceso, traducido a tres idiomas, y se mejoran las opciones de búsqueda del catálogo posibilitando la consulta por materias.

El PADICAT captura el patrimonio digital de Catalunya siguiendo el modelo híbrido de captura que hemos comentado. Por una parte captura masivamente las webs de acceso público del dominio .cat, y por otro lado establece acuerdos con instituciones y organismos para proceder a la captura de sus páginas web. Además, realiza campañas puntuales de captura vinculadas a acontecimientos relevantes en la vida del país, como las campañas electorales. A modo de ejemplo, podemos citar algunas de las monografías, que llevan por título Elecciones europeas (2009), Elecciones municipales (2007), Folk-rock (2008) o Museos de Cataluña.

El socio tecnológico indispensable en el proyecto es el Centre de Serveis Científics i Acadèmics de Catalunya (CESCA), que ha desarrollado y diseñado algunas aplicaciones para mejorar el acceso y la recuperación de los recursos digitales depositados en el PADICAT. Estos recursos se capturan mediante el software *Heritrix*, desarrollado por *Internet Archive* y las Bibliotecas Nacionales Nórdicas. Y como sistema de gestión documental para asignar metadatos se usa el programa de código abierto *Web Curator Tool*, creación conjunta de la Biblioteca Británica y la Biblioteca Nacional de Nueva Zelanda. Ambas soluciones son software libre que gozan de un gran predicamento entre la comunidad de bibliotecas que archivan webs.

La consolidación de esta infraestructura ha permitido al PADICAT alcanzar unos niveles de captura que ellos mismos han cuantificado en su web institucional:

-Compilación semestral de 30.000 recursos del dominio .cat.

-Compilación semestral de 550 recursos de las 450 entidades con las que se ha firmado un convenio de colaboración.

-Compilación semestral de los 800 recursos procedentes de recomendaciones de los usuarios.

-Compilación única de 1.000 recursos de colecciones monográficas.

-Compilación diaria de una parte substancial de 30 publicaciones seriadas en línea.

Es una tarea muy valiosa que a fecha de hoy ha conseguido compilar un total de 13 terabytes de información que reflejan una parte importante del patrimonio digital catalán.

El PADICAT es, por lo tanto, el primer proyecto de archivo web que tiene lugar en España. Hasta el 2007 no se crea Ondarenet, el archivo electrónico del patrimonio digital vasco, y hasta el 2009 la Biblioteca Nacional de España no encarga al *Internet Archive* las primeras capturas del dominio .es. En su corta existencia, sólo siete años y medio, el PADICAT se ha consolidado como un proyecto líder, que ha sabido construirse una red local de contactos que le garantice el apoyo institucional, y que también se ha relacionado con archivos digitales internacionales a través de su participación en el *International Internet Preservation Consortium*, donde ha podido enriquecerse mediante el intercambio de experiencias y conocimientos del resto de instituciones de *web archiving* del mundo.

Durante la elaboración del presente trabajo nos pareció oportuno poder contar con la opinión de algún miembro del equipo del PADICAT, que no ofreciera su visión experta sobre la

propia institución y sobre algunas cuestiones como la preservación digital, la ley de depósito legal o la colaboración de las empresas e instituciones con el PADICAT. Nos dirigimos al señor Ciro Llueca, que es el coordinador del PADICAT, además de profesor asociado de la Facultad de Biblioteconomía y Documentación de la Universidad de Barcelona. El señor Llueca muy amablemente se ofreció a colaborar y a contestar el breve cuestionario que reproducimos a continuación:

Santi Lopera: ¿Podría explicar brevemente cuál es el proceso de firma de un convenio de colaboración entre una empresa o un particular y el PADICAT para que éste lleve a cabo una captura periódica de su web?

Ciro Llueca: Después de identificar la empresa o institución con la que queremos establecer un convenio, les comunicamos nuestro sistema de funcionamiento, los beneficios del PADICAT, y tramitamos un modelo de convenio (http://www.padicat.cat/ca/collabora-i-participa/estableix-un-conveni). Si las negociaciones prosperan se firma por duplicado el convenio. A partir de ese momento pasan tres meses hasta que se realiza la primera captura completa, y con posterioridad a esa fecha la captura es semestral.

SL: ¿Cuáles son las reacciones más frecuentes por parte de las instituciones que son contactadas por el PADICAT para capturar su web? ¿Se da el caso en que es la empresa o particular quien se dirige al PADICAT para pedir el archivo de su web?

CL: Depende de la naturaleza de la empresa o institución, y de su conocimiento sobre la Biblioteca de Catalunya. Algunos Ayuntamientos cierran favorablemente el expediente en treinta días (abrir el expediente, pasar le convenio por comisiones informativas y pleno, tramitar el convenio firmado), otros tardan años en hacer las mismas gestiones independientemente del tamaño del consistorio o de los recursos que tengan. El más rápido de todos, por ejemplo, fue el Ayuntamiento de Palafrugell (23.000 habitantes). Tenemos otros, mucho más importantes por población, que todavía lo están gestionando. En todo caso, las administraciones públicas suelen entender el rol de la Biblioteca de Catalunya (BC) con mayor facilidad que algunas empresas, que pueden no concebir que el servicio de la BC no tiene un coste directo para ellas (me refiero, lógicamente, de lo que pagan con sus impuestos). Recuerdo que Aigües de Barcelona, o Mango, por poner ejemplos de grandes empresas, tardaron poquísimo en firmar una vez informadas del proceso. Otras, del mismo

tamaño o más pequeñas, han podido tardar más, a menudo porque no hemos sido capaces de explicar muy sintéticamente cuál es el objetivo del PADICAT.

SL: En líneas generales, ¿cree usted que las empresas e instituciones entienden la importancia de la preservación digital o es necesaria una fuerte labor pedagógica por parte del PADICAT?

CL: Como he leído recientemente, hay dos tipos de soluciones para las empresas: las vitaminas (mejoran el sistema de funcionamiento, a medio y largo plazo) y las aspirinas (resuelven el mal de cabeza existente). La preservación digital, tal y como la entendemos mayoritariamente aporta beneficios a medio y largo plazo (vitamina). A menudo las empresas están centradas en el día a día, y no pueden –por recursos, por estrategia- pensar en cuestiones futuras como las que proponemos (acceso permanente a sus páginas web). Por lo tanto la pedagogía es imprescindible.

SL: ¿Tienen datos de las visitas y consultas al fondo del PADICAT?

CL: 6.000 visitas el año 2012.

SL: ¿Se tiene conocimiento de la existencia de trabajos de investigación efectuados a partir del fondo del PADICAT?

CL: Sí, esencialmente de márquetin y comunicación y ciencia política, también de escuelas de diseño. Y sobre todo de preservación digital, procedentes de lso estudios del Grado de Información y Documentación. A menudo también nos han contactado estudiantes del Máster de Archivística de la UAB para pedir algún dato concreto o alguna cuestión relacionada con la política del PADICAT.

SL: ¿Tienen pensado desarrollar (o ya se están aplicando) estrategias de difusión del fondo del PADICAT y de estimulación de su consulta por parte de la comunidad investigadora y el público en general?

CL: Trabajamos en dos líneas. La primera es la difusión en entornos profesionales de instituciones de la memoria (bibliotecas, archivos, museos), publicando artículos en revistas científicas y profesionales, presentaciones de congresos, etc. También vamos a universidades (ciencia política) a explicar las colecciones monográficas de las campañas electorales y implicamos a los profesores en la selección de recursos. La segunda es la difusión entre el público generalista, por medio de notas de prensa, presencia en la Vikipèdia, las redes sociales, etc.

SL: ¿Ha afectado la crisis económica al desarrollo del PADICAT? ¿Se han buscado formas alternativas de financiación o acuerdos de colaboración con otras instituciones, como por ejemplo el otro archivo digital de la península: ONDARENET?

CL: Desde el principio del proyecto se implicaron otras instituciones: CESCA, por la parte técnica; y la fundación puntCAT, para la difusión del proyecto y la captura coordinada del dominio .cat. Además se han establecido diversas acciones de colaboración con universidades (ESMUC para la selección de música folk-rock en línea, por ejemplo). El PADICAT, por otro lado, está en la cartera de servicios de la Biblioteca de Catalunya para los proyectos de patrocinio que se inician de forma habitual.

SL: El *Internet Archive* mediante su servicio *Archive-it* establece acuerdos con diferentes instituciones (universidades, archivos, museos, etc.) para que estas desarrollen sus propias colecciones de páginas web capturadas, e incluso lleva a cabo planes de formación como el *K-12 Web Archiving* que pretende difundir la importancia de la captura web. ¿El futuro del archivo web pasa por concienciar a la ciudadanía de que las responsabilidades de la preservación digital son un deber compartido? Más allá de las diferencias entre le PADICAT i el *Internet Archive* ¿se podrían llevar a cabo proyectos similares en el país?

CL: Sí, es un buen modelo. Sin duda el camino pasa por la cooperación.

SL: Estos días estamos viendo como la Biblioteca Británica está realizando una gran campaña de archivo web posibilitada por un cambio reciente en la ley de depósito legal. ¿Podría comentar brevemente si nuestra normativa favorece o dificulta la práctica del *web archiving*?

CL: Con sinceridad, pienso que la práctica del *web archiving* va mucho más allá de la regulación legal. De hecho, en otros países (Holanda, Suiza) ni existe la ley del depósito legal, y el compromiso con la preservación por parte de las instituciones es muy elevado, en el sector analógico y en el digital. Quiero decir, sin una política o estrategia, y sin recursos la ley no sirve para nada.

Dicho esto, la ley española es un buen texto legal. La versión de 1971 estaba claramente desfasada, pero la nueva ley (23/2011, del 29 de julio, que entró en vigor en enero del 2012) va por le buen camino. Pero hace falta que se desplegué vía real decreto para que sea efectiva. En eso estamos trabajando las diversas bibliotecas nacionales que formamos parte del grupo de trabajo de preservación digital del Consejo Español de Cooperación Bibliotecaria.

La ley española bebe, en cuestiones de *web archiving*, de un magnífico texto legal: la ley danesa (http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html). La ley británica, de hecho, existe desde el 2003.

SL: ¿Cuáles son los objetivos y retos del PADICAT para los próximos años?

CL: La misión del PADICAT es recoger, conservar y difundir le patrimonio digital de Catalunya nacido en Internet.

El sistema se basa en la aplicación de una serie de programas informáticos que permiten la captura, el almacenamiento, la organización, la preservación y el acceso permanente a las páginas web publicadas en Internet.

Sus objetivos son:

- Compilar masivamente el dominio .cat.

-Impulsar el depósito sistemático de la producción web de las entidades y las empresas de Cataluña.

- Promover líneas de investigación procesando de manera monográfica los recursos de eventos de la vida pública catalana, como por ejemplo campañas electorales en Internet, el fenómeno de la música en línea, los museos en Internet.

Después de unas etapas de nacimiento (2005-2006), crecimiento (2007-2008) y consolidación (2009-2011), a partir del 2012 se persigue sistematizar la capacidad de crecimiento con al logro de incorporar anualmente unas 75.700 versiones de aproximadamente 32.000 páginas web, procedentes de:

- Compilación semestral de 30.000 recursos del dominio .cat.

- Compilación semestral de 550 recursos de las 450 entidades con las que se ha llegado a un convenio de cooperación.

- Compilación semestral de los 800 recursos procedentes de recomendaciones de los usuarios.

- Compilación única de 1.000 recursos de colecciones monográficas.

- Compilación diaria de una parte substancial de 30 publicaciones seriadas en línea.

A estos logros concretos hay que añadir cuatro ejes permanentes de trabajo:

- Definición de las estrategias de preservación digital para el patrimonio nacido a Internet. PADICAT proporciona radiografías periódicas de la web catalana; detecta los formatos que experimentan a corto plazo problemas de ilegibilidad; identifica los lenguajes más usados, etc.

- Impulso a líneas de investigación a partir de la creación de colecciones monográficas que cuentan con la implicación de expertos de cada materia.

- Creación y mantenimiento de la hemeroteca digital en Internet, con la captura sistematizada de publicaciones digitales en serie. Actualmente, una muestra representativa en cuanto a tipos y contenidos, seleccionando las nacidas digitales, sin equivalente analógico.

- Cooperación con otros archivos web y depósitos de preservación digital, de bibliotecas, archivos y museos, para dar una respuesta eficiente a los retos de preservación digital y acceso a los recursos depositados.

SL: Los archiveros gestionamos, custodiamos y preservamos documentos de archivo, cuya definición oficial es la siguiente: “Es la información registrada con independencia de las características físicas, e intelectuales, producida o recibida y conservada por una organización o una persona en el desarrollo de sus actividades”. ¿Cree que los archiveros tenemos que empezar a pensar en las páginas web como documentos de archivo? ¿Cree posible un acercamiento a la web des de la metodología archivística?

CL: Me cuesta ver diferencias entre los profesionales que trabajan en bibliotecas y archivos y otros servicios de información, más allá de una procedencia formativa que puede ser diversa y que, en la práctica, casi no diferencia la praxis profesional. En muchos países archiveros y bibliotecarios trabajan de manera conjunta para garantizar la preservación del patrimonio, sea analógico o digital. En Cataluña nos ha costado más, pero confío que las nuevas generaciones de archiveros y bibliotecarios se centren en trabajar para el servicio público y no en defensas gremiales del siglo pasado. Es necesario sumar esfuerzos, y en Cataluña, en ese sentido, no hemos hecho todos los deberes que teníamos que hacer.

PARTE II EL *WEB ARCHIVING* Y LA ARCHIVÍSTICA

8 ¿Es la web un documento de archivo?

Al principio del presente trabajo advertíamos de que el término “archivo web” es una aportación de los bibliotecarios y profesionales de la información que, como hemos visto, han sido los principales protagonistas del origen y desarrollo de esta técnica. Comentábamos que, aunque utiliza la misma palabra que da nombre a nuestra disciplina, es una técnica totalmente distinta a nuestra profesión. Pero a tenor de lo que hemos visto hasta ahora ¿realmente es el *web archiving* una propuesta extraña para la archivística? ¿Capturar, gestionar, conservar y hacer accesibles evidencias documentales no son tareas de los archiveros?

La principal pregunta que tenemos que abordar frontalmente para saber si el archivo web debe ser una de las preocupaciones de los archiveros es la siguiente ¿son las páginas web documentos de archivo? Revisemos la definición que la *Norma de Descripció Archivística de Catalunya (NODAC)* da de documento de archivo:

“Es la información registrada, **con independencia de sus características físicas e intelectuales**, producida o recibida y conservada por una organización o una persona en el desarrollo de sus actividades”. (NODAC, pág. 180)

Si hacemos caso de esta definición es evidente que las páginas web pueden ser documentos de archivo y, en consecuencia, los archiveros deberíamos intentar aproximarnos a su tratamiento desde nuestra propia metodología. Una cuestión nada fácil si tenemos en cuenta el carácter dinámico de las páginas web, que son objetos digitales complejos que cambian con menor o mayor periodicidad. Pero una cuestión, al fin y al cabo, que tenemos que afrontar desde nuestra experiencia y conocimiento teórico si queremos ser protagonistas de la gestión documental del siglo XXI.

¿Son las bibliotecas o los archivos los que deben archivar las páginas web? Hasta ahora son las bibliotecas las que lo hacen, pocos archivos trabajan en este campo a nivel mundial. Y no conocemos ninguno que lo haga en nuestro país. Para mí, la respuesta adecuada a este interrogante es que las páginas web pueden ser gestionadas por bibliotecas y también pro archivos. Cada profesional le dará un tratamiento diferente, porque su objeto de estudio y su metodología son diferentes. Las bibliotecas buscan preservar la producción digital del país, atesorar las expresiones culturales y científicas de la sociedad para que las generaciones futuras las puedan estudiar y conocer.

Pero los archivos deben preocuparse las evidencias documentales que las páginas web contienen. Los archivos no deben dirigir su mirada a Internet como si fuera un inmenso océano en el que lanzar las redes para ver qué capturan. No es así como trabajan los archivos. Los archivos lo son en función de la organización para la que trabajan, no tratan colecciones, sino fondos documentales que, como sabemos, vienen determinados por un productor. ¿Tiene este productor una web corporativa? ¿La utiliza como un escaparate para comunicar al mundo su imagen? ¿O la utiliza como un instrumento a través del cual relacionarse con sus clientes y proveedores? ¿Se de intercambio de información en esta página web? ¿Es esta información evidencia de las actividades desarrolladas por la organización?

Este es el proceso que considero que deben seguir los archivos con el *web archiving*. Deben tomar esta metodología, descomponerla en piezas pequeñas y volver a montarla bajo criterios archivísticos. Como hemos visto es una disciplina que durante diecisiete años ha evolucionado bajo unos criterios determinados entre los cuales no se contaban los archivísticos. Pero el archivo web tiene mucho que ofrecer a la profesión, sus instrumentos serán muy útiles cuando, tarde o temprano, los archiveros debamos abordar la cuestión de las páginas web, porque la evolución de la comunicación pasa por Internet de forma ineludible. Por lo tanto, dado que es un camino que tenemos que recorrer, mejor no retrasarse más.

9 ¿Por qué los archivos de Cataluña no archivan las páginas web?

Un análisis profundo sobre la historia de la archivística en Cataluña supera las intenciones del presente trabajo, y los conocimientos de su autor, pero sí querría intentar apuntar algunas posibles explicaciones que respondan a la pregunta de por qué los archiveros no archivamos las páginas web. Ante todo es un intento de responderme a mí mismo este interrogante que, hasta ahora, no ha sido planteado teniendo en cuenta que nos amparan criterios científicos y profesionales para poder desarrollar esta labor con el convencimiento de que forma parte de nuestro campo de actuación, y de que es responsabilidad del archivero y gestor documental tratar archivísticamente las evidencias documentales que contienen las páginas web.

No hay una única respuesta que contesta esta pregunta, sino que será la combinación de una serie de factores la que podría explicar esta situación. Por un lado podemos decir que Cataluña tiene una gran tradición archivística pero no ha visto profesionalizada esta disciplina hasta una fecha muy reciente. Se ha hecho mucho trabajo desde la llegada de la democracia para desarrollar un sistema archivístico cohesionado en todo el territorio, aprobando normativas y leyes para garantizar la integridad dle patrimonio documental. Se han establecido pautas para actuar con la documentación, de han fijado criterios de restauración y conservación, se han definido perfiles profesionales, se ha implantado un sistema de gestión documental que contempla archivos de gestión, archivos intermedios y archivos históricos. Se han creado instituciones capitales como el *Arxiu Nacional de Catalunya*, que vela por la conservación del patrimonio documental catalán, o la *Comissió Nacional d'Accés Avaluació i Tria Documental* que ha de definir el período de retención y el régimen general de acceso de los documentos. Se ha hecho mucho trabajo y muy bueno en un período relativamente corto de tiempo, y esta tarea se tiene que reconocer, pero el espíritu crítico será el que nos haga avanzar y mejorar. Y existen síntomas que nos indican que la profesión archivística en Cataluña no está plenamente normalizada: no existe un Colegio Profesional de archiveros (sí una *Associació d'Arxivers i Gestors Documentals*), no existe una carrera universitaria de archivística homologable fuera de Cataluña (sí la *Escola Superior d'Arxivística i Gestió Documental*), la consideración que la administración tiene de la archivística es menor y así se demuestra cuando la arrincona en el ámbito cultural en lugar de reconocerle su carácter transversal que verdaderamente le permitiría incidir en la gestión documental de las instituciones públicas.

No pierdo de vista el tema del presente trabajo cuando hablo de todo esto. El carácter lateral que la Administración otorga a la archivística es probablemente la razón por la cual hace

bastantes años tenemos la asignatura pendiente de la administración electrónica. Y de rebote, claro, la del archivo web que como hemos visto, no es imaginable sin un sistema previo de gestión de documentos electrónicos. Los gestores documentales tenemos muy poca incidencia en las soluciones tecnológicas que hasta ahora se han buscado en la Administración, y muy frecuentemente se quiere construir la casa por el tejado evitando los pasos intermedios. Es sabido que para implantar un sistema de gestión de documentos electrónicos se deben automatizar los procesos de trabajo. Para automatizarlos, previamente, se deben diagramar para entender los flujos de trabajo y la producción documental que generan. Y antes incluso, y aquí está la pieza clave de todo bajo nuestro punto de vista, para poder diagramar los procesos es preciso entender cómo trabaja una organización hasta el más mínimo detalle. Es una ardua labor de hormiguita, que requiere de mucho tiempo y que implica dialogar con los trabajadores implicados en cada uno de los procedimientos para comprender bien de qué manera se trabaja. Esta no es una tarea que genere resultados a corto plazo, y es muy costosa en tiempo y esfuerzos. ¿Cuántas organizaciones, corporaciones, agencias y entidades públicas (o empresas privadas) conocemos que tengan diagramados y automatizados todos sus procedimientos?

Cualquier solución tecnológica que se busque deberá ser adaptada a la manera de trabajar de la institución en la que se implante. Y este conocimiento previo tenemos que aportarlo los archiveros y gestores documentales, que conocemos como nadie las funciones y actividades del organismo para el que trabajamos. Considero que hasta que este paso no se de la administración electrónica no acabará de funcionar, de la misma manera que un edificio no puede sostenerse sin cimientos sólidos. Y hasta que la Administración no se de cuenta, o los gestores documentales le hagamos darse cuenta, que le camino adecuado pasa por dotar a los servicios de archivo de más capacidad de acción y protagonismo en el funcionamiento de las entidades no aprobaremos esta asignatura.

Otras razones que explican el hecho de que los archivos no archiven las páginas web también las podemos buscar dentro de la misma profesión. Quizás por esta adscripción equivocada de la archivística al ámbito cultural, o por el elevado tanto por ciento de archiveros que provienen de los estudios de Humanidades (yo mismo soy licenciado en Historia), exista en la profesión cierta resistencia a las nuevas tecnologías. He tenido la ocasión de comprobarlo entre los compañeros de la ESAGED y, recientemente, ha vuelto a salir este tema de discusión en la lista de distribución Arxiforum[14]. De forma esquemática se podría hacer una división entre una tendencia más historicista, que valora la importancia cultural e histórica de los documentos y entiende le archivo como un edificio en el que

[14] Arxiforum és el primer fòrum electrònic de discussió sobre la teoria i la pràctica de l'arxivística que des de 1998 administra l'Associació d'Arxivers – Gestors de Documents de Catalunya.

custodiar los documentos de gran significación cultural; y por otro lado, otra tendencia que defiende la importancia de la gestión documental como un conjunto de principios y metodologías necesario para le buen funcionamiento de una organización.

Bajo mi parecer, el que piense que la figura del archivero debe tener poca relación con la tecnología se equivoca rotundamente. Con toda seguridad los primeros archiveros de la historia eran escribanos, y los escribanos eran expertos en la tecnología de la escritura, que no era algo que estuviera al alcance de cualquiera. Conocían perfectamente el proceso de preparación de las pieles para confeccionar pergaminos, conocían la composición de las tintas que se utilizaban, la variedad que existía i el tiempo necesario para secarse, así como los utensilios más adecuados para escribir. Por lo tanto, un requisito imprescindible de la profesión de archivero y gestor documental debería ser el de conocer la tecnología que genera los documentos que se custodian. Naturalmente que el estallido de las nuevas tecnologías de la información y la comunicación no lo pone nada fácil, pero en lugar de resistencia como mínimo debería haber entre los archiveros un interés por aprender el funcionamiento de los nuevos medios. Falta apertura de miras en este sentido, y falta formación específica que ayude a los archiveros y gestores documentales a afrontar retos como los de la preservación digital o el del *web archiving*.

10 Incorporación del archivo web a las tareas de archivo

Abordar el reto de desarrollar tareas de archivo web desde los archivos no es algo diferente al de hacer frente a la gestión de documentos electrónicos. Prácticamente es lo mismo, es ir un paso más allá y otorgar a las páginas web la importancia que ya están adquiriendo en nuestra sociedad. Las páginas web deberían ser tratadas como documentos electrónicos –y, de hecho, lo son- en la gestión de los cuales el tratamiento de los metadatos es un factor absolutamente clave.

Un primer paso ineludible debería ser el de legar a un consenso, en el seno de cada organización, sobre qué partes de la web se consideran documentos de archivo. Puede que la totalidad de la web, una parte o solamente los objetos digitales que contiene. Es necesario llevar a cabo esta reflexión teórica para establecer cuáles serán las unidades básicas con las que trabajaremos más tarde. Con seguridad las alternativas serán diversas, a tenor de la complejidad que las páginas web han ido adquiriendo con la evolución tecnológica. Han pasado de ser, en un primer momento, simples imágenes fijas hasta convertirse en complejas bases de datos interrelacionados. Será cuestión de que cada institución defina, en función del uso que le dé a su propia web, qué partes de esta constituyen evidencias de sus actividades.

Este es un debate que los Archivos Nacionales de Australia tuvieron hace más de diez años y como resultado publicaron unas recomendaciones para gestionar documentos de archivo basados en la web[15]. Estas recomendaciones contemplan dos estrategias diferentes:

-*Object-driven approach*. Que podríamos traducir como aproximación basada en el objeto. Más apropiada para webs estáticas, que funcionan como almacén de objetos digitales. Estas páginas web se podrían capturar en su totalidad conservando una panorámica global (*snapshot*) i, con posterioridad, llevar a cabo un seguimiento automatizado de los cambios que se den en el tiempo: nuevos objetos añadidos, objetos modificados, etc.

-*Event-driven approach*. Aproximación basada en el evento. Es un tratamiento apropiado para páginas web diseñadas para interactuar con el usuario. Estas web acostumbran a trabajar con bases de datos de perfiles de usuario, mecanismos de búsqueda, acciones programadas y otras funcionalidades. Por lo tanto, sería necesario poder registrar todas las

[15] National Archives of Australia

acciones que un usuario efectúa en el entorno web, capturar su dirección IP o dominio, el perfil de usuario, fecha y hora de las acciones, etc.

Una vez definida esta cuestión, otro paso a considerar sería el momento en el que estos documentos basados en la web deben ser capturados e introducidos en el sistema de gestión documental de la organización. La respuesta debería darse en relación a la función de tales documentos. Si van ligados a un expediente o trámite dentro del cual constituyen una evidencia, o tienen importancia como piezas informativas. La propia tramitación debería determinar el momento en que deberían capturarse. Por otro lado, si las capturas son panorámicas y se realizan para testimoniar la evolución de la web, se podrían automatizar a intervalos de tiempo regulares. En función del nivel de actividad de la web podrían ser semanales, mensuales, trimestrales, etc.

Una vez que estos documentos basados en la web han entrado en el sistema de gestión documental su tratamiento sería el mismo que el del resto de documentos electrónicos. Serían clasificados según el cuadro de clasificación de la institución, y se les aplicarían las políticas de disposición, seguridad y acceso pertinentes.

El requisito previo, por lo tanto, para que una organización pueda desarrollar tareas de archivo web de sus páginas web es que disponga de un sistema de gestión documental orientado al tratamiento de documentos electrónicos, con todas las herramientas que eso requiere: cuadro de clasificación, calendario de disposición, esquemas de metadatos, cuadro de seguridad y acceso, etc.

11 Escenarios posibles

Si hacemos el ejercicio de imaginar de qué manera los archivos públicos de Catalunya pueden comenzar a integrar tareas de *web archiving* en su trabajo diario se nos ocurre que podrían darse dos posibles escenarios de futuro.

Red descentralizada de archivos web

El escenario más probable podría ser aquel en el que los archivos con más capacidad (presupuestaria, tecnológica, etc.) iniciasen sus propios proyectos de archivo web. Serían archivos que ya disponen de un sistema de gestión de documentos electrónicos implementados, y que querrían ir un paso más allá y empezar a capturar, tratar y custodiar las páginas web de su institución.

Sería deseable que estas instituciones publicasen un reglamento propio que defina cuáles son los documentos web que deben capturar y preservar. Incorporar estas nuevas responsabilidades no debería suponer un gran inconveniente para un servicio de archivo que ya trabaje con documentos electrónicos. Se necesitaría un software de captura y gestión de páginas web, el conocimiento necesario para utilizarlo y el suficiente espacio de almacenamiento digital. Pero para una institución que ya cuente con un repositorio digital seguro ampliar la capacidad de almacenamiento no debe entrañar un problema, y más si tenemos en cuenta que éste no debe ser muy grande dado que la porción de internet que se quiere capturar es mínima. Son las bibliotecas nacionales las que quieren conservar grandes cantidades de información, los archivos tendrían suficiente con las evidencias documentales de las webs de su institución

Este modelo descentralizado ampliaría la brecha digital que actualmente ya existe en el Sistema de Archivos de Cataluña, donde se dan situaciones dispares de archivos que sólo trabajan en papel y otros que se han podido dotar de instrumentos más avanzados que les permiten gestionar documentos electrónicos.

Archivo web central, bajo el paraguas institucional de l'ANC

El otro escenario posible sería la creación de un repositorio institucional centralizado que diera servicio a todo el sistema de Archivos de Cataluña. Igualmente sería imprescindible la aprobación y publicación de un reglamento que marcara los principios y las pautas de actuación de la captura web. O mejor que un reglamento sería una normativa fruto del consenso de comisiones de expertos y grupos de trabajo, con voluntad globalizadora.

Cada archivo del SAC capturaría sus propios recursos web pero estos serían almacenados en un repositorio centralizado compartido con toda la red de archivos. Un gestor documental implementaría los metadatos necesarios que identificarían la organización generadora de las capturas, y facilitaría la posterior recuperación de la información. Esto obligaría a iniciar un proceso de equiparación tecnológica de aquellos archivos que hasta el momento no hayan tenido la oportunidad de dotarde se estos medios. Sería necesaria la instalación del software necesario para la captura en cada archivo, y garantizar los recursos tecnológicos necesarios para que las capturas lleguen al repositorio central a través de un canal seguro. Tecnológicamente es algo relativamente sencillo actualmente y no se necesitaría maquinaria especial, ya que existen mecanismos para crear canales seguros de intercambio de información por Internet utilizando sistemas de encriptación específicos.

Por lo tanto, la inversión económica sería contenida teniendo en cuenta que existe software libre de captura web, y que los archivos sólo necesitarían un ordenador con conexión a Internet. La parte fuerte de la inversión iría destinada a la creación del gran repositorio central, con todos los requisitos tecnológicos necesarios, hardware y software específicos, seguridad de las instalaciones, personal, etc.

En este modelo se garantizaría una mayor cohesión entre los miembros del Sistema de Archivos de Cataluña, estableciendo unos mínimos protocolos de actuación y estandarizando unas prácticas comunes emanadas de un órgano central director, que coordine los esfuerzos e impulse las políticas definidoras de la práctica archivística.

12 Conclusiones

A lo largo de la historia de la humanidad se dan momentos en los cuales las innovaciones tecnológicas impactan de tal manera sobre el comportamiento de las sociedades que es difícil comprender la dimensión de su importancia. Frecuentemente se requiere de perspectiva histórica para poder evaluar de forma más acertada no sólo de qué manera el avance tecnológico revolucionó la vida material del ser humano, sino, lo que es incluso más importante, de qué manera le hizo ampliar su horizonte de pensamiento. Cómo pulveriza los esquemas mentales y permite a las comunidades humanas subir un escalón más en su etapa evolutiva. Algunos ejemplos son el descubrimiento del fuego en la noche de los tiempos, la invención de la rueda, el desarrollo del astrolabio, la creación del telescopio, la invención de la imprenta, la fabricación de la máquina de vapor, etc.

Todas estas tecnologías produjeron cambios tan incomparables que las sociedades humanas necesitaron un tiempo para asimilarlos y poder adaptarse al nuevo mundo de posibilidades que abrían. De la misma manera, nuestra sociedad del siglo XXI está adaptándose a la irrupción de las nuevas tecnologías de la información y la comunicación y especialmente al nacimiento de Internet, que lo está cambiando todo. Y aún no podemos alcanzar a comprender de qué manera tan profunda está removiendo las raíces del mundo contemporáneo. Las posibilidades de la comunicación instantánea con cualquier parte del mundo, las enormes cantidades de información disponible para cualquier persona con acceso a la red, el enorme poder organizativo de las redes sociales, etc.

No caeremos en la ingenuidad de creer que Internet es la solución a todos los problemas de la humanidad; es un medio, y el ser humano es el responsable del uso que le dé. Eugeny Morozov ya ha advertido del poderoso instrumento de control que para os regímenes autoritarios supone Internet, y del peligro del *ciberutopismo*, que cree que universalizando le acceso a Internet se diluirán las desigualdades sociales. Todo esto forma parte del período de adaptación al nuevo paradigma tecnológico, estudiar las posibilidades que ofrece, y entender los peligros que puede entrañar.

Las sociedades que más rápidamente y mejor se adapten a los cambios tecnológicos tendrán mucho ganado. A lo largo del presente trabajo hemos visto como una profesión como la de bibliotecario ha reaccionado a la irrupción de Internet y ha conseguido entender sus responsabilidades en esta nueva situación. Ha reorientado sus prioridades y ha desarrollado el archivo web, que es la tecnología que les permite continuar ejerciendo sus obligaciones de custodios de la cultura y el patrimonio editorial nacional en el nuevo paradigma tecnológico.

Los archiveros tenemos que entender que los sistemas de producción de documentos han cambiado y, sin dejar a un lado el patrimonio documental analógico, que se debe continuar tratando, custodiando y explotando, tenemos que adaptarnos a la gestión del documento electrónico. Podemos encontrar elementos útiles en el *web archiving* que nos ayuden a llevar a cabo este proceso de adaptación, dado que cada vez más los documentos se crean en las páginas web. Estudiar y analizar la tecnología y los procedimientos que los bibliotecarios y ingenieros de la información han desarrollado para capturar las webs nos puede ayudar a construir nuestra propia manera de capturar los documentos que se crean e intercambian en un entorno web. Bajo nuestra propia metodología y nuestros principios teóricos, para mantener estas evidencias documentales íntegras, auténticas, recuperables y utilizables. No existen razonamientos científicos que impidan a los archiveros tratar las páginas web como documentos de archivo.

Este paso que debemos dar los archiveros como profesión es una fase más de la adaptación a la gestión de documentos electrónicos. Para poder ir a buscar documentos a las páginas web para integrarlos en el sistema de gestión documental de nuestras organizaciones, antes, aunque resulte evidente decirlo, tenemos que haber implantado plenamente un sistema propio de gestión de documentos electrónicos. Y este es el punto en e que nos encontramos hoy día. Tenemos el conocimiento y sabemos qué herramientas necesitamos. Para tratar la documentación analógica teníamos suficiente con un cuadro de clasificación y algunos instrumentos de descripción que nos ayudaran a recuperar los documentos. Pero para asumir plenamente la gestión de documentos electrónicos tenemos que diagramar los procesos y automatizarlos, es decir, reflexionar profundamente alrededor de cómo y para qué trabajamos.

Y hasta que este paso no se haya dado en todas las piezas del Sistema de Archivos de Cataluña no será posible la creación de un archivo web centralizado bajo el paraguas institucional del Arxiu Nacional de Catalunya, como apuntábamos en el punto 4 de la segunda parte del trabajo. Porque no es una cuestión de pura inversión tecnológica, sino de reconfigurar la manera de trabajar de cada institución. Y los archiveros y gestores documentales deberíamos de tener la responsabilidad de liderar e impulsar estos cambios organizativos. Por eso es imprescindible que, por un lado, los archiveros y gestores documentales mejoren sus competencias tecnológicas, para entender con mayor profundidad los nuevos sistemas de producción documental digitales. Y, por el otro, poder disfrutar de un mayor reconocimiento por parte de las cúpulas de las organizaciones que les otorgue más poderes y capacidades decisorias, para propiciar estos cambios en el sino de las instituciones desde el conocimiento experto que solo el archivero y gestor documental puede tener.

13 Glosario

Open-source: término que se ha traducido al castellano como código abierto. Hace referencia al software distribuido y desarrollado de forma libre. Su característica es que da acceso al código en el que ha sido programado y esto puede permitir a los usuarios que lo adapten a sus necesidades y colaboren en su desarrollo.

Backup: es una copia de seguridad que se hace de unos datos con la finalidad de poder recuperarla en el caso de pérdida de la información original.

Ciberutopismo: es un término creado por el pensador y escritor Eugeny Morozov en su obre *The Net Delusion: The Dark Side of Internet Freedom* (*La ilusión de Internet: el lado oscuro de la libertad de Internet)* que hace referencia a la creencia de que la comunicación por Internet favorece a los oprimidos más que a los opresores.

Cookies: es el término que designa una pequeña parte de información que envía un sitio web y almacena el navegador del usuario, de manera que el sitio web puede consultar la actividad previa del usuario.

URL: del inglés *Uniform Recurs Locator*. Es una cadena de caracteres que informa al navegador de la máquina donde está el recurso al que hace referencia; el protocolo que hay que utilizar para obtener este recurso; y la manera como el servidor web encontrará el recurso.

Web crawler: (también conocido como araña web) es un software que inspecciona las páginas web de forma automatizada. Su función acostumbra a ser la de realizar una copia de las páginas web, que posteriormente serán indexadas para facilitar su recuperación.

14 Bibliografía i fuentes

Ayre, Catherine; Muir, Adrienne. "The right to preserve: the rights issues of digital preservation". *D–lib magazine*. Volumen 10, núm. 3. (Marzo de 2004). En línea <http://www.dlib.org/dlib/march04/ayre/03ayre.html>

Aguillo, Isidro F. "Internet Invisible o Infranet: Definición, clasificación y evaluación". *Jornadas Españolas de documentación* (7enes, 2000, Bilbao). En línea <http://www.fesabid.org/repositorio/jornadas-espanolas-de-documentacion/actas-de-las-vii-jornadas-espanolas-de-documentacion>

Brügger Niels. *Archiving Websites. General considerations and strategies*. Aarhus: The Centre of Internet Research, 2005. En línea < http://cfi.au.dk/publications/books/>

Cordón, José Antonio. "El Depósito legal y los recursos digitales en línea". *Las bibliotecas nacionales del siglo XXI*. Valencia: Biblioteca Valenciana, 2006. P. 97-114. En línea <http://eprints.rclis.org/15036/>

Crook, Edgar. "*Web archiving in a web 2.0 world*". *Australian Library and Information Association (ALIA) Biennial Conference: Dreaming 2008*. (6ª, 2008, Alice Springs). En línea < http://conferences.alia.org.au/alia2008/papers/>

Departament de Cultura, Subdirecció General d'Arxius. *Norma de Descripció Arxivística de Catalunya (NODAC) 2007*. Generalitat de Catalunya, 2007, Barcelona.

Gomes, Daniel; Miranda, Joao; Costa, Daniel. "A survey of web archiving iniciatives". *International Conference of Theory and Practice of Digital Libraries* (15avas, 2011, Berlín). En línea < http://sobre.arquivo.pt/about-the-archive/a-survey-on-web-archiving-initiatives>

ICA/ATOM–Committee on electronic records. *Guide for managing electronic records from an archival perspective*. ICA, febrero 1997, París. En línea <http://www.ica.org/11878/digital-recordkeeping-programme-resources/.html>

Kahle, Brewster. "Editors' Interview: The Internet Archive." *RLG DigiNew. Volum* 6, núm. 3. (15 junio 2002). En línea: <http://www.rlg.org/preserv/diginews/diginews6-3.html#interview>

Kuny, Terry. "A Digital Dark Ages? Challenges in the Preservation of Electronic Information". *Annual Conferences of IFLA* (63avas, 1997, Copenhagen). En línea: <http://archive.ifla.org/IV/ifla63/63kuny1.pdf>

Llueca, Ciro. "Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales". *BiD: textos universitarios de biblioteconomía i documentación*. Núm. 15 (diciembre de 2005). En línea <http://www.ub.edu/bid/15lluec1.htm>

Llueca, Ciro. "Archivando la Web, el proyecto Padicat (Patrimonio Digital de Cataluña)". *El profesional de la información*. Volumen 15, núm. 6 (2006), p. 473-478. En línea <http://eprints.rclis.org/8399/>

Llueca, Ciro; Cócera-Saló, Daniel. **"***PADICAT: realitat i reptes de 3 anys de l'arxiu web de Catalunya"*, 2008. *Jornades Catalanes d'Informació i Documentació* (11avas, 22-23 mayo del 2008, Barcelona). En línea <http://eprints.rclis.org/11626/>

Llueca, Ciro; Cócera-Saló, Daniel; Torres, Natalia; Suades Méndez, Gerard; De-la-Vega-Sivera, Ricard. "*CAT (Curator Archiving Tool): millorant l'accés als arxius web". International Internet Preservation Consortium meeting* (2010 B, Viena). En línea <http://eprints.rclis.org/14902/>

Llueca, Ciro; Cócera-Saló, Daniel;Torres, Natalia; Suades-Méndez, Gerard; De-la-Vega-Sivera, Ricard. **"**El PADICAT, l'experiència catalana en l'arxiu d'Internet". *Lligall*, n. 31 (2010 A). P. 143-161. En línea <http://eprints.rclis.org/16246/>

Llueca, Ciro; Cócera-Saló, Daniel; Torres, Natalia; Suades Méndez, Gerard; De-la-Vega-Sivera, Ricard. "*PADICAT, el archivo de Internet"*, 2011. *Jornadas Españolas de Documentación*, (12avas, 2011 A, Málaga). En línea <http://eprints.rclis.org/15761/>

Llueca, Ciro; Cócera-Saló, Daniel; Torres, Natalia; Suades-Méndez, Gerard; De-la-Vega-Sivera, Ricard. **"**A ritmo de tweet: archivando elecciones 2.0". *El profesional de la información*. Volumen 20, núm. 3 (2011 B). P. 309-314. En línea <http://eprints.rclis.org/15764/>

Llueca, Ciro; Reoyo-Tudó, Sandra. "Repositoris digitals: disseny i implementació per a biblioteques, arxius i museus". Presentación en el curso 47/12 del COBDC (octubre de 2012). En línea <http://eprints.rclis.org/17923/>

Lyman, Peter. "Archiving the World Wide Web". *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington DC, Council on Library and Information Resources and Library of Congress, 2002. P. 38-51. En línea <http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42>

Masanes, J. "Towards continuous Web archiving: First results and an agenda for the future". *D-Lib Magazine* Volumen 8, núm. 12 (diciembre de 2002). En línea <http://www.dlib.org/dlib/december02/masanes/12masanes.html>

Masanes, J. *Web archiving*. Masanes, J. (ed.) 2006.

Mohr, Gordon [et al.]. "An introduction to Heritrix: an open source archival quality web crawler". *International web archiving workshop*, (4as, 2004, Bath). En línea <http://bit.ly/U2xbSI>

National Archives of Australia, *Archiving web resources: Guidelines for keeping records of web-based activity in the Commonwealth Government*. 2001.

Pennock, M; Kelly, B. "Archiving web sites resources: a records management view". *International World Wide Web Conference* (15avas, 2006, Edinburg). En línea: <http://opus.bath.ac.uk/424/>

Pulgar Vernalte, Francisca; Marcos Maciá, Sonia. *"Ondarenet: el archivo del patrimonio digital vasco"*. *Jornadas de Gestión de la Información* (10as, 2008, Madrid). En línea <http://eprints.rclis.org/12553/>

Pulgar Vernalte, Francisca; Marcos Maciá, Sonia. "*Capturing the Basque Web"*. *Libraries in the Digital Age LIDA*, (2as, 2009, Dubrovnik i Zadar). En línea <http://eprints.rclis.org/13164/>

Soler, Joan. *Del bit al logos. Preservar documents electrònics a l'Administració local*. Diputació de Barcelona, octubre de 2003, Barcelona.

Soler, Joan. "Algoritmes diplomàtics (IX): les mil i una denominacions de document electrònic". *Diplomàtica.cat*, 9 de octubre de 2012. <http://diplomaticapuntcat.blogspot.com.es/2012/10/algoritmes-diplomatics-ix-les-mil-i-una.html>

Térmens, Miquel. "La sostenibilitat econòmica i tècnica dels repositoris de preservació digital". *Lligall*, n. 31 (2010). P. 44-62. En línea:

<http://arxivers.com/publicacions/revista-lligall/edicions-lligall/cat_view/14-revista-lligall/61-lligall-31.html>

UNESCO. *Directrices para la preservación del patrimonio digital*. UNESCO, 2003, Canberra. En línea <http://bit.ly/U2whWh>

15 Anexos

Listado de miembros del International Internet Preservation Consortium. No constituyen todas las iniciativas de archivo web existentes pero sí son las más destacadas.

Institución - Nombre del archivo		Categoría
Biblioteca Alexandrina		Fundación sin ánimo de lucro
URL	http://www.bibalex.org/isis/frontend/archive/archive_web.aspx	

Biblioteca Nacional de Francia		Biblioteca Nacional
URL	http://www.bnf.fr/fr/collections_et_services/collections_departements.html	

Bibliotecas de la Universidad de Columbia		Universidad/Investigación
URL	http://www.archive-it.org/organizations/304	

Biblioteca y Archivos de Canadá - Archivo Web del Gobierno de Canadá		Biblioteca Nacional
URL	http://www.collectionscanada.gc.ca/	

Biblioteca de Harvard - Servicio de archivo web de colecciones de la Universidad de Harvard		Universidad/Investigación
URL	http://wax.lib.harvard.edu/collections/home.do	

Biblioteca Nacional y Universitaria de Croacia - Archivo Web Croata		Biblioteca Nacional
URL	http://haw.nsk.hr/	

Instituto Nacional del Audiovisual (Francia)		Biblioteca Nacional
URL	http://www.ina.fr/	

Internet Archive		Fundación sin ánimo de lucro
URL	http://archive.org/index.php	

Internet Memory Foundation		Fundación sin ánimo de lucro
URL	http://internetmemory.org/en/	

Biblioteca Nacional y Universitaria de Islandia - Archivo Web Islandés	Biblioteca Nacional
URL	http://vefsafn.is/

Biblioteca Nacional de Finlandia - Archivo Web Finlandés	Biblioteca Nacional
URL	http://webarchive.nationallibrary.fi/

Biblioteca Nacional de Suecia - Kulturarw3	Biblioteca Nacional
URL	http://www.kb.se/om/projekt/Svenska-webbsidor---Kulturarw3/

Biblioteca del Congreso de los Estados Unidos - Archivo Web de la Biblioteca del Congreso	Biblioteca Nacional
URL	http://www.loc.gov/lcwa

Royal Library and The State and University Library (Aarhus) - Netarkivet.dk	Biblioteca Nacional
URL	http://netarkivet.dk/in-english/

Biblioteca Nacional de Noruega - Archivo Web de Noruega	Biblioteca Nacional
URL	http://www.nb.no/fag/nasjonalbibliotekets-samling/nettdokumenter2

Biblioteca Nacional de Nueva Zelanda - Archivo Web de Nueva Zelanda	Biblioteca Nacional
URL	http://natlib.govt.nz/collections/a-z/new-zealand-web-archive

Biblioteca Nacional de Corea - OASIS	Biblioteca Nacional
URL	http://www.oasis.go.kr/ctrlu?cmd=main

Biblioteca Nacional de Australia - PANDORA	Biblioteca Nacional
URL	http://pandora.nla.gov.au/

Biblioteca de Catalunya - PADICAT	Biblioteca Nacional
URL	http://www.padicat.cat/

Biblioteca Nacional de España - Archivo Web Español	Biblioteca Nacional

URL	http://www.bne.es/es/Inicio/index.html

Biblioteca Nacional y Universitaria de Eslovenia - Archivo Web de Eslovenia	Biblioteca Nacional
URL	http://www.nuk.uni-lj.si/

Los Archivos Nacionales del Reino Unido - Archivo Web del Gobierno del Reino Unido	Biblioteca Nacional
URL	http://www.nationalarchives.gov.uk/webarchive/

Biblioteca Británica - Archivo Web del Reino Unido	Biblioteca Nacional
URL	http://www.webarchive.org.uk/ukwa/

Biblioteca Nacional Diet – Proyecto de Archivo Web de Japón	Biblioteca Nacional
URL	http://warp.da.ndl.go.jp/search/

Biblioteca Digital de California - Servicio de Archivo Web	Universidad/Investigación
URL	http://webarchives.cdlib.org/

Biblioteca Nacional de los Países Bajos - Archivo Web de los Países Bajos	Biblioteca Nacional
URL	http://www.kb.nl/

Biblioteca Nacional de la República Checa - Archivo de la Web Checa	Biblioteca Nacional
URL	http://webarchiv.cz/

Biblioteca Nacional de Austria - Archivo Web de Austria	Biblioteca Nacional
URL	http://www.onb.ac.at/ev/about/webarchive.htm

Biblioteca Nacional de Suiza - Archivo Web de Suiza	Biblioteca Nacional
URL	https://www.e-helvetica.nb.admin.ch/pages/main.jsf