

EL *WEB ARCHIVING* I L'ARXIVÍSTICA

SANTI LOPERA LOPERA

Director del treball: Joan Soler

Any d'elaboració: 2013

Treball de recerca del Màster d'Arxivística i Gestió de Documents

Escola Superior d'Arxivística i Gestió de Documents

Col·lecció Treballs fi de màster i de postgrau

Com citar aquest article: Lopera Lopera, Santi. (2013) *El web archiving i l'arxivística*. Treball de recerca del Màster d'Arxivística i Gestió de Documents de l'Escola Superior d'Arxivística i Gestió de Documents. (Treballs fi de Màster i de postgrau). [Http://...](http://...) (consultat el ...)



Aquesta obra està subjecta a llicència Creative Commons Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya (<http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.ca>). Es permet la reproducció total o parcial i la comunicació pública de l'obra, sempre que no sigui amb finalitats comercials, i sempre que es reconegui l'autoria de l'obra original. No es permet la creació d'obres derivades.

Resum

L'arxivament del web és una disciplina que té 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.

Título: El *web archiving* y la archivística

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.

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.

SUMARI

Introducció.....	7
PART I. WEB ARCHIVING.....	9
1 Què és el <i>web archiving</i> ?	10
2 Quan es comença a arxivar el web?	12
3 Per què arxivar el web?.....	13
4 Com s'arxiva el web?	15
4.1. <i>Crawlers</i>	15
4.2. Models de captura	15
4.3. Polítiques.....	16
5 Entrebancs del <i>web archiving</i>	18
5.1. Dificultats tècniques.....	18
5.2. Dificultats legals.....	21
5.3. Dificultats d'organització	21
5.4. Dificultats econòmiques	22
6 Qui arxiva el web?.....	24
6.1. Biblioteques nacionals	24
6.2. Projectes supranacionals	26
6.3. Universitats i organismes d'investigació.....	28
6.4. Empreses.....	29
7 L'arxivament del web a Catalunya: el PADICAT	30
PART II. EL WEB ARCHIVING I L'ARXIVÍSTICA.....	38
8 És el web un document d'arxiu?.....	39
9 Per què els arxius de Catalunya no arxiven els webs?	41
10 Incorporació de l'arxivament web a les tasques d'arxiu.....	44
11 Escenaris possibles	46
12 Conclusions	48

13	Glossari	50
14	Bibliografia i fonts	51
15	Annexos	55

Introducció

La voluntat d'aquest treball és la de cridar l'atenció de la comunitat arxivística del país al voltant d'una de les branques de la preservació digital que més expansió es preveu que ha de tenir durant els pròxims anys, l'arxivament del web, i que fins ara és una pràctica inèdita als nostres arxius. Internet és un instrument que des de fa anys està experimentant una expansió que no té frenada, i no només és el mitjà per excel·lència a través del qual les persones intercanvien tot tipus de béns, d'informació, de dades, sinó que també és un sistema de producció de documents de primera magnitud.

Aquests documents que viatgen pel ciberespai, de màquina a màquina, de servidor a servidor, de pàgina web a pàgina web, no acostumen a ser tractats arxivísticament, amb la consegüent pèrdua de contextualització que això comporta, i de garanties de fiabilitat, integritat, i autenticitat que la disciplina arxivística és capaç de garantir. No es pot dir, a juny de 2013, que la gestió dels documents electrònics sigui quelcom nou que s'acaba d'inventar, i malgrat això és una pràctica que no està plenament implantada als arxius de Catalunya.

El tractament de les pàgines web, la seva captura, gestió i preservació és una faceta més de la gestió de documents electrònics que ha estat oblidada durant massa temps i potser ha arribat el moment de començar a pensar la manera com els arxivers i gestors documentals podem afrontar-la. Pensem, per exemple, en les seues electròniques que l'Administració desenvolupa per relacionar-se amb els ciutadans. Aquestes pàgines web contenen documents amb valor jurídic, generadors de drets i obligacions, i potser fóra bo aplicar l'arxivament del web a les seues electròniques per garantir una millor gestió de la seva informació.

El treball es divideix en dues parts, en la primera s'intenta donar resposta de forma breu i fugint de tecnicismes, a qüestions com ara en què consisteix l'arxivament del web, quan es va començar a aplicar, de quina manera es duu a terme, o quines són les principals institucions que el practiquen avui dia. Intentem entendre també la peculiaritat de les pàgines web, i com la seva natura dinàmica i efímera dificulta el seu tractament des d'un punt de vista arxivístic. Aquests reptes requereixen noves respostes tecnològiques que han de ser desenvolupades sota els criteris de la gestió documental per poder garantir un tractament correcte de les unitats documentals que es creen, s'intercanvien o s'emmagatzemen a les pàgines web.

La primera part del treball es tanca amb un breu repàs als principals projectes d'arxivament web del món, realitzats per institucions líders que marquen la pauta del desenvolupament tecnològic i teòric, però que no ho fan sota criteris arxivístics ja que en la seva major part són biblioteques nacionals. Comentem també, i de forma destacada, el cas del PADICAT

(Patrimoni Digital de Catalunya), de la Biblioteca Nacional de Catalunya, un projecte de referència internacional desenvolupat al país.

La segona part del treball abandona el caràcter expositiu i intenta relacionar la pràctica de l'arxivament del web, creada i desenvolupada des de la biblioteconomia i les ciències de la informació, amb la disciplina arxivística. És, per tant, un apartat més arriscat en el qual l'autor intenta apropiarse al *web archiving* des d'una perspectiva arxivística, destacant la importància que la gestió documental comenci a veure les pàgines web com un mitjà més que s'ha de tenir present, i sobre el qual els arxivers han d'estendre les seves responsabilitats donat que serà el principal mitjà d'intercanvi de documents en un futur proper.

Intentem donar una explicació, segurament incompleta donada la complexitat de la qüestió però meditada i sincera, de per què l'arxivament del web és una pràctica que no es realitza als arxius de Catalunya i imaginem quins escenaris possibles poden donar-se en el futur.

El present treball és, per tant, una aportació humil i modesta que vol recordar a la comunitat arxivística catalana la importància de la preservació digital, i en concret de l'arxivament del web, i la necessitat cada cop més imperiosa de donar respostes als reptes que la gestió de documents electrònics planteja. Qüestió avui dia indefugible si no volem perdre el cavall de la innovació tecnològica, i que l'arxivística i la gestió documental es converteixin en disciplines oblidades i arraconades per causa de la seva incapacitat d'aportar solucions eficaces als problemes que el nou entorn tecnològic genera.

PART I. *WEB ARCHIVING*

1 Què és el *web archiving*?

El *web archiving* o arxivament del web són un conjunt de tècniques que tenen com a finalitat la captura de les pàgines web. És una pràctica que s'emmarca dins de la preservació digital, que és la disciplina que es preocupa per la conservació a llarg termini dels objectes digitals, independentment del seu format, del programari amb el que van ser creats o del maquinari que els conté. La constant i imparable evolució de la informàtica ha donat un enorme ventall de sistemes per crear i processar la informació, i gran part d'aquest programari i maquinari ha quedat obsolet amb el pas del temps i ha sigut substituït per un altre. La preservació digital treballa per evitar la pèrdua d'aquesta informació i desenvolupa diferents tècniques (emulació, migració, captura de metadades, etc.) per a què aquests objectes digitals siguin accessibles per a les societats del futur.

És important adonar-se de la velocitat amb la qual els documents digitals poden quedar desfasats i per copsar-ho d'una manera ràpida només cal donar un cop d'ull al llistat de formats que recull PRONOM¹, la base de dades de formats creada pel departament de preservació digital dels Arxius Nacionals del Regne Unit, que en recull més de 900. A la Wikipedia, en canvi, han compilat més de 3.500² formats diferents, tot i que encara és un llistat incomplet. La principal característica dels documents digitals és que necessiten d'una màquina per ser llegits³, i si no disposem de màquines amb capacitat per interpretar la codificació amb la qual van ser creats els documents, aquesta informació s'ha perdut per sempre.

Per entendre les dificultats de conservació de les pàgines web hem de tenir en compte que aquestes són una estructura de múltiples documents, de diferents formats i tipus, interrelacionats entre ells (bases de dades, formularis, documents de text, etc.) susceptible de patir problemes d'obsolescència si no rep el degut tractament. Altres factors que determinen la vida de les pàgines web són els relacionats amb el servidor que els dona allotjament. Aquest servidor pot patir problemes tècnics, o extingir-se el contracte del servei de manteniment d'un web i comportar la desaparició d'una part o de la totalitat de la pàgina.

Per tant, la preservació digital és un dels reptes més importants amb què es troba l'arxivística moderna, que no només ha de garantir que els documents electrònics siguin autèntics, fiables, íntegres i complets, sinó també accessibles. Fins ara, per garantir l'accessibilitat dels documents al llarg de la història era suficient amb el fet que l'arxiver els

¹ <http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=listReport>

² Col·laboradors 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](http://en.wikipedia.org/w/index.php?title=List_of_file_formats_(alphabetical)&oldid=537895641)>

³ ICA/ATOM pàg. 24.

guardés en un lloc ben vigilat i tingués cura que no els malmetessin els insectes o els rosegadors. Avui dia guardar els documents electrònics en un repositori d'accés controlat no assegura la seva conservació pels problemes d'obsolescència que hem comentat, i es requereix de tot un seguit d'accions preventives que la preservació digital s'encarrega d'estudiar.

L'arxivament del web és, per tant, la branca de la preservació digital que s'encarrega de l'estudi i desenvolupament de tècniques, programari i maquinari per a la captura de les pàgines web i la conservació d'aquestes a llarg termini.

Com veurem en el següent punt, les primeres en començar a capturar pàgines web van ser les biblioteques nacionals i, per tant, el terme *web archiving* prové del món de la biblioteconomia i de les ciències de la informació, no pas de l'arxivística. És important que aquesta qüestió etimològica quedi clara des de bon principi.

Aquest arxivament del web no segueix la metodologia arxivística ni té com a objecte d'estudi ni de treball els documents d'arxiu, tal i com són entesos en la nostra disciplina. Molt pocs arxius arreu del món practiquen l'arxivament del web, i es dóna per tant la paradoxa que no són els arxiviers qui arxiven el web si no els bibliotecaris i professionals de la informació. Pot semblar una qüestió menor però creiem important assenyalar aquest fet per evitar confusions, així com plantejar des d'aquí que, en ares d'un llenguatge més clar, potser seria millor trobar altra expressió que defugui l'ambigüitat del terme arxivament del web. I més encara si tenim en compte que l'arxivística d'una banda i la biblioteconomia i la documentació d'una altra són disciplines diferents però amb algunes fronteres compartides, que han fet esforços per delimitar el seu camp d'estudi i deixar ben clares les seves competències.

2 Quan es comença a arxivar el web?

Les pàgines web es comencen a capturar quan es popularitza el seu ús i creació de la mà de l'expansió d'Internet. Oficialment es considera que el naixement d'Internet es produeix l'1 de gener de 1983⁴, amb la primera xarxa de llarg abast basada en tecnologia TCP/IP que va posar en marxa la *National Science Foundation* dels Estats Units. Però no és fins la dècada dels noranta que es popularitza el seu ús a partir de dues fites cabdals. D'una banda, el naixement l'any 1991 del projecte *World Wide Web*. I d'altra banda, el treball de l'enginyer informàtic britànic Timothy Berners Lee, que el 1993 va desenvolupar el llenguatge de marques HTML i la creació del protocol d'intercanvi HTTP.

És en aquell moment quan comencen a proliferar les pàgines web, que qualsevol ciutadà d'arreu del món pot generar, i a mesura que va augmentant l'accés a Internet i, consegüentment, el contingut de la xarxa, es desperta la preocupació sobre com gestionar aquesta quantitat de coneixement. Les biblioteques nacionals, que tradicionalment havien atresorat la totalitat de publicacions del propi país mitjançant la gestió del dipòsit legal, posen en marxa iniciatives per continuar desenvolupant aquesta missió pròpia en el món digital.

D'aquesta manera és durant l'any 1996 quan es configuren tres projectes pioners, *Kulturarw3* de la Biblioteca Nacional de Suècia, *Pandora* de la Biblioteca Nacional d' Austràlia, i l'*Internet Archive*, que es funda a San Francisco (Califòrnia) però que no és una iniciativa nacional sinó que és una entitat sense ànim de lucre amb vocació internacional.

Posteriorment, les biblioteques nacionals d'altres països van anar creant els seus propis projectes d'arxivament del web i els repositoris digitals on conservar aquesta informació. Com per exemple l'arxiu web de Nova Zelanda, l'arxiu web de la Biblioteca del Congrés dels Estats Units, l'arxiu web de Noruega o, ja a casa nostra, el PADICAT, el Patrimoni Digital de Catalunya, creat per la Biblioteca de Catalunya el 2005.

Per tant, la història del *web archiving* és molt curta, uns disset anys des de la posada en marxa dels projectes pioners, però molt activa donat que en un període de temps tan curt han nascut més de 40 projectes d'arxivament del web en 26 països diferents⁵ que han capturat més 7.500 terabits d'informació.

⁴ <http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet>

⁵ Gomes, Daniel

3 Per què arxivar el web?

Obligació legal

Com hem vist anteriorment, l'impuls inicial d'arxivar les pàgines web prové de les biblioteques nacionals, que tenen com un dels pilars bàsics de la seva missió institucional recopilar la producció bibliogràfica, discogràfica, etc., del país, conservar-la, i fer-la accessible als ciutadans. El llistat d'obres enregistrament obligatori al Dipòsit Legal és molt variat⁶ i va des de partitures musicals fins als cartells publicitaris que podem veure a les carreteres. Les obres digitals que es produeixen a Internet també són d'interès del Dipòsit Legal, i és obligació de les biblioteques nacionals fer-se'n càrrec. L'arxivament del web ofereix a les biblioteques la possibilitat de capturar aquest patrimoni del país ja sigui per donar compliment a l'obligació legal en si, o per desenvolupar la vocació de defensa i promoció de la cultura que tradicionalment ha caracteritzat aquestes institucions.

Investigació

Amb el desenvolupament d'Internet s'ha produït una ràpida transformació en la manera com la nostra societat es relaciona i comunica. L'expansió de l'accés a la xarxa ha fet que Internet es converteixi en el mitjà principal en el qual es donen tot tipus d'intercanvis, des dels purament econòmics (transaccions bancàries, compra i venda de productes, etc.) fins als més clarament informatius, laborals, socials, etc. Internet ha transformat la manera en la qual gran part de la humanitat es comunica i és un reflex directe de l'activitat de les comunitats que hi participen. Aquest aspecte ha estat reforçat per l'esclat de les xarxes socials, veritables finestres on els individus exposen i transmeten la seva vida en temps real.

Tot aquest enorme volum d'informació és la matèria primera amb la qual els historiadors de la societat del segle XXI hauran de treballar. I per a què això tingui lloc serà imprescindible la conservació de les pàgines web que testimonien aquesta vitalitat. Però també els investigadors de les més diverses disciplines requeriran de l'arxivament del web per la seva posterior consulta i estudi; sociòlegs, antropòlegs, economistes, especialistes en màrqueting, periodistes, historiadors de l'art, i un llarg etcètera tindran en un futur —i tenen ja avui dia— en Internet el seu camp d'estudi.

⁶ <http://www.bnc.cat/Professionals/Diposit-legal#quinesobres>

Memòria universal del coneixement

D'altra banda, també cal destacar que existeix una línia de pensament intrínsecament lligada al naixement i desenvolupament d'Internet que defensa una sèrie de principis com el de la neutralitat a la xarxa, els drets digitals (dret d'accés a Internet, llibertat d'expressió i d'informació a la xarxa, etc.), que considera la xarxa un bé universal. És una filosofia molt propera als principis de la Wikipedia, per exemple, segons la qual grans comunitats d'individus treballen de forma desinteressada pel bé comú de la compilació del coneixement.

Des d'aquest punt de vista, Internet és una font d'informació i coneixement universal a la qual tothom ha de poder accedir, incloses les generacions futures. I l'arxivament del web es presenta com una de les grans estratègies de preservació digital que pot garantir que aquest patrimoni de la humanitat no es perdi. Amb aquest esperit va néixer l'arxiu digital més gran avui dia l'*Internet Archive*, que dóna accés lliure a tot el seu dipòsit.

Aquests tres conceptes són arguments poderosos que justifiquen l'exercici del *web archiving* però no tenen perquè ser els únics. Un altre més concret és l'interès que cada productor de pàgina web ha de tenir a conservar la seva informació publicada en línia pel seu propi ús, ja siguin empreses, universitats, institucions, administracions, o particulars.

4 Com s'arxiva el web?

La finalitat d'aquest treball no és tractar en profunditat l'aspecte tècnic del *web archiving* però sí voldríem fer cinc cèntims sobre la manera com es duu a terme per tal de proporcionar una idea general de la seva metodologia a nivell intel·lectual, sobretot, per el lector que s'apropa per primera vegada a aquesta disciplina.

4.1. Crawlers

La captura de les pàgines webs es realitza mitjançant un programari conegut amb el nom de *web crawler*. És un programari automatitzat (*internet bot*) que té com a finalitat buscar, capturar i indexar les url's de les pàgines web. Per tant, el contingut d'Internet que es captura depèn de les instruccions que se li donen als *crawlers*.

De la gran quantitat de *crawlers* existents podríem destacar *Heritrix*, un programari de codi lliure desenvolupat per l'*Internet Archive* i les Biblioteques Nacionals Nòrdiques. Algunes de les organitzacions que el fan servir són institucions que lideren el panorama de l'arxivament web com la Biblioteca Nacional d'Austràlia, la Biblioteca Nacional de França o la Biblioteca Britànica.

Un altre *crawler* que seria bo destacar és *HTTrack*. Aquest va ser desenvolupat per Xavier Roche i permet descarregar les pàgines web capturades en el propi ordinador. Un altre captador web que podríem mencionar és *Wget*, en aquest cas molt lligat al sistema operatiu GNU⁷.

Cal destacar també que la proliferació d'empreses de serveis de captura web ha donat naixement a tota una àmplia gama de captadors que no entrarem a comentar aquí, ja que ens allargaríem potser massa. Direm només que tres dels més coneguts són *Archive-It*⁸, *Archivethe.net*⁹ i *Hanzo archives*¹⁰.

4.2. Models de captura

En funció dels criteris que es facin servir per dur a terme la captura i l'arxivament de les pàgines web podem establir tres enfocaments diferents:

⁷ <http://www.gnu.org/gnu/gnu.html>

⁸ <http://www.archive-it.org/>

⁹ <http://archivethe.net/fr/>

¹⁰ <http://www.hanzoarchives.com/>

-Model integral o exhaustiu: pretén la captura automatitzada de totes les pàgines web sota un criteri general, ja sigui lingüístic (per exemple, totes les pàgines en català), d'ubicació dels servidors (per exemple, les pàgines allotjades en servidors ubicats a Catalunya), de domini (per exemple, totes les pàgines amb domini “.cat), etc.

Aquest model assegura la compilació de gran quantitat d'informació però no pot accedir a la Internet invisible, que és aquella part d'Internet que es genera de forma dinàmica quan un usuari fa una consulta, o que està protegida per contrasenya (en el punt 5.1. en parlem més). Aquesta manera massiva de capturar pàgines web no permet un tractament acurat de la informació recopilada i, per tant, presenta més limitacions en el moment de ser posteriorment recuperada.

-Model selectiu: segueix uns criteris que acostumen a ser temàtics i, abans d'efectuar la captura, l'entitat arriba a acords amb els editors i productors dels recursos web. És per tant una tasca més lenta que requereix contactar amb productors web, negociar les condicions i establir acords però, en canvi, la informació capturada es pot tractar al detall i assegurar una explotació posterior més eficient. És un sistema costós, tenint en compte els recursos humans necessaris per dur-lo a terme.

-Model híbrid: consisteix en una combinació dels dos models anteriors. Molts projectes nacionals com per exemple el PADICAT l'apliquen, i intenten capturar el web de forma sistemàtica i també arribant a acords amb les institucions productores. Puntualment es fan captures selectives amb un motiu social o històric destacat que tindrà lloc en un període de temps molt concret, com ara la celebració d'uns jocs olímpics o bé un procés electoral.

4.3. Polítiques

L'arxivament del web no consisteix en posar a un robot a capturar pàgines web, sinó que abans d'això, i també amb posterioritat a la captura, hi ha tot un seguit de decisions que les organitzacions han de prendre. Són unes decisions amb les quals els arxivers estem familiaritzats ja que són força semblants a les normes de gestió documental que recomana la ISO-15489-1 i la ISO-15489-2. D'una forma molt esquemàtica podem dividir-les en els següents apartats:

Polítiques de captura

Cada institució estableix els paràmetres dels seus propis capturadors en funció de la missió i els objectius propis. Ha de tenir en compte la llei de propietat intel·lectual i la de continguts digitals del seu país, que li indicarà els límits de captura que no s'han d'excedir. Ha de tenir en compte el pressupost de la institució així com les solucions tècniques que farà servir per a la captura.

Política de gestió i accés de les col·leccions

Són totes aquelles decisions relatives al tractament de les pàgines web capturades: ordenació, classificació, indexació, etc. Serà necessari establir quins són els responsables del tractament i determinar si el contingut és d'accés obert o bé si té alguna restricció i, en aquest cas, identificar quins són els subjectes que poden accedir-hi.

Política de preservació

A través d'aquesta política ha de determinar-se el contingut que s'ha de conservar de manera permanent i el que ha de ser eliminat. En conseqüència s'ha de fixar el calendari d'eliminació per una part, i, per l'altre, les tècniques de preservació que es duran a terme (migració, emulació, etc.) en funció de les característiques tècniques de les col·leccions. Altre aspecte que s'ha de tenir en compte és el de la seguretat i manteniment del repositori digital en el qual les col·leccions es conserven, i els requisits que ha de complir per tal de garantir la seva confiabilitat.

5 Entrebancs del *web archiving*

Com a conseqüència de la pròpia naturalesa d'Internet, l'arxivament del web es troba amb tota una sèrie de problemes i limitacions de molt diversa índole que podríem comentar breument en quatre apartats.

5.1. Dificultats tècniques

Creixement exorbitant

En primer lloc, cal destacar que Internet és un espai que no para de créixer; és finit però està en constant transformació i canvi, i creix a un ritme superior a la velocitat amb la qual es captura el web. Només cal fer-se dues preguntes per copsar aquesta diferència de ritme: Quants agents fan créixer Internet? Quants agents capturen el web? Els segons mai podran igualar els primers i el dubte de si es podria fer un *backup*, una còpia de seguretat, de tot Internet queda així fàcilment resolt. Altra qüestió és si cal replicar la totalitat d'Internet, pels arxivers, que avaluen i conserven només una part dels documents que generen les institucions on treballen, no és necessari. Però algunes de les organitzacions que actualment arxiven el web, com hem vist, sí tendeixen a capturar una gran part d'Internet i han de fer front al repte d'aquesta gran expansió.

Natura dinàmica i caràcter efímer del web

D'altra banda, les pàgines web, com a objectes digitals, són força especials, ja que en la seva majoria no constitueixen un únic document electrònic sinó que acostumen a ser una arquitectura de múltiples documents, de diferents formats i tipus, interrelacionats entre ells. A més, hem de tenir en compte el caràcter dinàmic del web, que és la característica que el converteix finalment en un objecte tremendament complex.

Brügger¹¹ indica que el dinamisme del web es dóna de tres maneres diferents:

-Des del punt de vista de qui envia la informació:

- a) Dinàmica d'actualització. Un web creat pot ser modificat en qualsevol moment.
- b) Dinàmica de proliferació. En qualsevol moment es poden crear nous webs.

¹¹ Brügger, Niels. Capítol 3.

-Des del punt de vista dels elements textuais (ja sigui text, imatges, sons, etc.):

a) Dinàmica del moviment. Entre elements textuais i dintre dels mateixos, ja sigui amb enllaços que redireccionen al visitant, o descàrregues que s'executen quan es clica sobre un botó preconfigurat.

b) Dinàmica de la complexitat. A mesura que un web integra més i més elements interrelacionats entre ells i amb les accions que realitzarà el visitant, augmenta la seva complexitat.

-Des del punt de vista del recipient que conté la informació

a) Dinàmica de l'equipament. El web es visualitzarà de forma diferent en funció del navegador que l'executi o del dispositiu (telèfon mòbil, tauleta tàctil, ordinador personal, etc.)

b) Dinàmica de les accions. Les accions que un visitant fa sobre el web el poden personalitzar, per exemple mitjançant les *cookies*, i això fa que canviï el seu comportament en funció del visitant.

Un aspecte lligat al dinamisme del web és el seu caràcter efímer, les pàgines web en constant actualització eliminen part dels seus continguts amb elevada freqüència. Pensem en una pàgina web de notícies: quantes vegades al dia canvia la portada d'un diari digital? Quan es captura un web es pretén reproduir una petita part de la realitat que existeix a Internet en un repositori digital, però el contingut guardat deixarà de tenir una equivalència amb la realitat quan la pàgina es modifiqui, i només quedarà la captura com a testimoni d'un moment concret. Fins i tot es pot donar el cas que la captura enregistri una realitat que mai va tenir lloc. És el que passaria si una pàgina s'actualitza mentre s'està capturant. Seguint amb l'exemple del diari digital, la portada capturada barrejaria notícies d'abans i de després de l'actualització, i seria un híbrid que no reflectiria la realitat amb fidelitat.

Per tant, les pàgines web estan constantment canviant o desapareixent, i això vol dir que una enorme quantitat d'informació s'està perdent per a les generacions futures.

Internet profunda o *deep web*

Aquest terme fa referència a la part d'Internet que no poden indexar els motors de cerca habituals i que, per tant, està fora de l'accés de la majoria d'internautes. Tampoc els *crawlers* que fan servir els arxius digitals per capturar les pàgines web poden accedir a la Internet profunda.

Els principals motius pels quals els motors de cerca no poden accedir a la Internet profunda són tres:

- Els editors web que protegeixen les seves pàgines per evitar que aquestes siguin indexades, ja sigui amb contrasenyes, tallafocs o fitxers d'exclusió de robots.txt.
- La gran part dels recursos d'Internet que requereixen d'una consulta a la base de dades per poder accedir a la seva informació. Mentre aquesta no es faci, el contingut no es fa visible (per exemple: diccionaris o enciclopèdies).
- El fet que el contingut es trobi en un format no indexable.

Aguillo¹² estableix que la Internet profunda està formada per quatre grups de recursos digitals:

- Bases de dades bibliogràfiques: formades per catàlegs de biblioteca accessibles mitjançant una passarel·la web, altres bases de dades de referències bibliogràfiques i catàlegs de llibreries.
- Bases de dades alfanumèriques: totes les bases de dades que no tinguin caràcter bibliogràfic i altres recursos que requereixen d'una passarel·la web per a la seva consulta (enciclopèdies, diccionaris, etc.).
- Arxius i revistes electròniques: acostumen a ser recursos de gran qualitat als quals s'accedeix exclusivament amb una consulta prèvia.
- Fitxers no HTML o textuals: com per exemple l'Acrobat PDF.

Estàndards web i formats de software canviant

La vertiginosa evolució de la tecnologia produeix nous formats, nous programaris i noves maneres de concebre, dissenyar i implementar una pàgina web. Aquest factor dificulta la feina de l'arxivament del web, que ha d'anar adaptant-se constantment a aquests canvis i ha de fer evolucionar els seus propis sistemes de captura i emmagatzematge per adaptar-se a les noves situacions.

També l'evolució dels llenguatges de marques i dels estàndards web requereix d'una adaptació dels agents que capturen el web. Per motius com aquests és important la tasca de l'*International Internet Preservation Consortium*, que promou el desenvolupament i ús d'eines comunes, tècniques i estàndards que facilitin la creació d'arxius internacionals.

¹² Aguillo, Isidro F.

L'esperança del Web semàntic

El projecte de construir una web semàntica que faciliti la interoperabilitat dels sistemes informàtics i faciliti la interpretació de la informació per part de les màquines potser sigui molt útil per l'arxivament del web. Aquest projecte, ideat per Tim Berners-Lee, promou l'ús d'uns llenguatges de marques estàndards per descriure la informació *on line* de tal manera que les pàgines web siguin més comprensibles per les màquines i aquestes puguin operar amb el significat de la informació de forma automàtica.

Sens dubte, seria un gran avenç a l'hora de consultar els arxius web, ja que la connexió automàtica i significativa entre les dades donaria a l'usuari uns resultats molt més rics que la majoria de cercadors d'avui dia.

5.2. Dificultats legals

Un altre entrebanc que han de gestionar les entitats que practiquen l'arxivament del web és la difícil qüestió dels drets d'autor dels continguts d'Internet i el conflicte que pugui derivar-se de la protecció de dades personals. Davant d'això no és possible una estratègia estandarditzada degut al fet que en cada país existeix una reglamentació diferent sobre aquestes matèries i cal un enfocament diferent per a cada situació.

Les possibles estratègies que es poden desenvolupar al respecte són tres:

- Capturar i arxivar exclusivament webs lliures de drets d'autor.
- Desenvolupar polítiques efectives de gestió de drets.
- Capturar massivament però només permetre l'accés en funció del sol·licitant. És el que es fa, per exemple, a Dinamarca, on la llei de dipòsit legal permet des del 2005 capturar la part del web desitjada sense haver de demanar permís als propietaris. Per contra, per no vulnerar els drets, no es permet l'accés al públic general.

5.3. Dificultats d'organització

En aquest punt volíem comentar el fet de que Internet no està governada per cap entitat i creix i es desenvolupa de manera descentralitzada. Tot i que són molts els agents i *lobbies* que volen exercir el seu control mitjançant l'aprovació de lleis i reglaments, en realitat no podem dir que Internet sigui un espai que pertany a una sola entitat. El fet positiu que Internet no pertanyi a una sola entitat té el contrapunt negatiu, des del punt de vista de

l'arxivament del web, que els estàndards i les recomanacions que facilitarien la captura no es poden imposar i són responsabilitat de cada individu i de cada institució productora.

D'altra banda existeixen diferents maneres de fer front a l'arxivament del web depenent de la institució que el dugui a terme, ja siguin Biblioteques Nacionals, museus, centres d'història, galeries d'art, etc. Cadascú tindrà la seva pròpia estratègia i els seus propis objectius, però si es vol avançar en matèria de preservació digital del contingut *on line* és necessari reforçar la cooperació entre els agents, estalviar esforços i no repetir captures.

5.4. Dificultats econòmiques

El fet tècnic de capturar i arxivar una pàgina web no ha de ser una operació molt costosa econòmicament parlant. Existeix programari lliure de captura, i com a repositori podem fer servir el disc dur del nostre ordinador, qualsevol disc dur portàtil que tinguem o, fins i tot, l'espai que moltes empreses ofereixen al núvol de forma gratuïta. Així doncs, capturar el web de forma amateur és ràpid i barat.

Però, com hem vist, el sentit últim de l'arxivament del web és el de salvaguardar grans quantitats d'informació per evitar que es perdin, i aquesta és una tasca impossible de realitzar de manera individual i que requereix del lideratge de grans projectes de captura i emmagatzematge, i és aquí on la situació canvia radicalment i els costos econòmics es comencen a disparar. A banda d'un software de captura seria necessari un programari de gestió que ens possibilités treballar la informació, indexar-la i aplicar-li polítiques de gestió, d'accés, de preservació digital, etc. Hauríem de tenir un sistema per fer accessible les captures, amb una *interface* que permetés la consulta. Necessitaríem un equip de persones expertes en la matèria que gestionessin el programari de captura i el tractament posterior de la informació.

Seria del tot imprescindible comptar amb un maquinari especialment dissenyat per emmagatzemar grans quantitats d'informació, amb les mesures de seguretat suficients com per evitar atacs exteriors però que al mateix temps estigués *on line* per facilitar la consulta a distància. S'haurien d'implementar sistemes de còpies de seguretat periòdiques per assegurar la perdurabilitat de la informació, i polítiques de migracions de formats per garantir la preservació. Com podem imaginar els costos econòmics derivats de la posada en marxa i el manteniment de tota aquesta estructura tecnològica i humana són elevadíssims. Miquel Térmens¹³ conclou que l'evolució tecnològica comportarà un abaratiment dels costos derivats del maquinari i l'emmagatzematge, però que el veritable repte d'aquests grans

¹³ Térmens, Miquel

repositoris digitals massius no és el seu manteniment en el temps si no el fet de poder disposar del capital inicial suficient com per engegar el projecte.

Per tant, és clau una correcta planificació prèvia, i del tot recomanable, gairebé imprescindible, aliances entre organismes i institucions per poder portar a terme de forma coordinada i compartida projectes de captura web sòlids, robusts i duradors.

6 Qui arxiva el web?

Com hem vist amb anterioritat qui aplicava l'arxivament del web en la seva fase inicial eren les biblioteques nacionals, però durant els disset anys d'història d'aquesta tècnica els canvis tecnològics i organitzatius han fet que actualment trobem més actors amb un paper destacat dins d'aquest escenari.

Actualment existeixen més de seixanta projectes d'arxivament web a tot el món, seria molt extens parlar-ne de tots ells i estaria fora de les intencions d'aquest treball. Per tant, el que hem fet és establir quatre grans categories on classificar aquests actors, per comentar-les breument i destacar alguns exemples representatius.

6.1. Biblioteques nacionals

Són les institucions que primer van encarar la problemàtica de capturar les pàgines web i les que primer van desenvolupar solucions tecnològiques per poder engegar l'arxivament del web. El fet de ser institucions úniques i rellevants, amb un pes específic al seu país, ha fet que poguessin comptar amb pressupostos suficients com per endegar els seus respectius projectes. En molts casos són la única institució que se n'ocupa al país i això, d'alguna manera, les ha convertit en projectes nacionals i en actors de referència.

Els projectes de les biblioteques nacionals tenen com a denominador comú la cerca d'aquella part d'Internet produïda pel propi país o que d'alguna manera hi pertanyi tenint en compte criteris històrics, socials o culturals. Amb tot, no és una tasca senzilla determinar quina part d'Internet pertany a un determinat país quan la xarxa és un espai vast desproveït de fronteres.

Una de les pioneres és la **Biblioteca Nacional de Suècia**, l'arxiu web de la qual rep el nom de **Kulturarw3**. No és estrany que aquesta institució fos de les primeres en preocupar-se per aquesta qüestió, i és que Suècia ha sigut molt precoç en aquesta matèria i ja el 1661 va encarregar a la Biblioteca Reial les tasques de col·leccionar totes les publicacions impreses del país. Per tant, aquesta preocupació d'atresorar la pròpia producció cultural els ve de molt antic.

El 1996 inauguren el projecte del *Kulturarw3* i comencen a plantejar-se qüestions com la de definir quina part d'Internet es podria considerar que pertany a Suècia. Mitjançant una estratègia exhaustiva, capturen tots els servidors amb domini suec (.se), però també els servidors amb qualsevol altre domini però enregistrats a Suècia (ja sigui amb una adreça postal sueca o amb un número de telèfon suec). També realitzen la captura de diaris on-line, i en capturen diàriament uns 140. Malgrat tot, la legislació del país no permet l'accés al

contingut de *Kulturaw3* per Internet, i només es pot consultar el fons des de la mateixa Biblioteca Nacional de Suècia.

Altra gran iniciativa pionera va ser la de **PANDORA**, l'arxiu web de la **Biblioteca Nacional d' Austràlia** que va encetar-se l'any 1996. Val a dir que no és l'únic projecte d'aquestes característiques al continent (també es poden mencionar *Our Digital Island* de Tasmania, i *Territory Stories* de la Biblioteca de l'Estat de *Northern Territory*) però sí és el més important i un dels que va començar a definir aquesta disciplina tal i com es coneix ara.

PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) ha construït una xarxa col·laborativa formada per nou peces, les biblioteques dels sis estats continentals més el Memorial Australià de la Guerra, l'Arxiu Nacional d'Imatge i So, i l'Institut Australià d'Estudis Aborígens i dels Habitants de l'Estret de Torres. PANDORA és un exemple d'arxiu web selectiu que no pretén capturar tota la producció nacional, sinó que fa una selecció i tria entre aquells recursos que considera més valuosos i dels que creu que han de ser preservats. Cadascú dels socis membres té les seves pròpies prioritats de captura, però tots treballen des d'aquesta estratègia selectiva i arxiven materials relatius a la vida social, cultural, política i intel·lectual dels australians. Això inclou blogs, pàgines web estatals, d'organitzacions, col·leccions de diaris i captures adreçades a enregistrar l'impacte d'un esdeveniment determinat.

Han desenvolupat el seu propi programari de gestió, PANDAS, que els permet gestionar les col·leccions de manera acurada, indexant, catalogant i controlant l'accés dels continguts. Aquesta manera de treballar tan detallada és molt lenta i els impossibilita capturar tots els recursos que voldrien preservar. Per evitar que aquests es perdin contracten els serveis de l'*Internet Archive* que és capaç, amb el programari *Heritrix*, de capturar un enorme espai web en molt poc temps, això sí sense el detall que proporciona PANDAS.

L'accés al contingut de PANDORA és totalment lliure per qualsevol persona amb connexió a la xarxa, tret d'una petita part (aproximadament el 2%) que gairebé en la seva totalitat es pot consultar només des dels ordinadors de la mateixa biblioteca.

L'últim projecte nacional del que parlarem és l'**Arxiu Web del Regne Unit**, que és un exemple de la capacitat de coordinació de múltiples institucions, en la seva majoria biblioteques nacionals, per construir un gran repositori digital representatiu de la importància històrica, social i cultural del país. L'Arxiu Web del Regne Unit va ser constituït el 2004 per la **Biblioteca Britànica** i al llarg de la seva trajectòria diverses institucions hi han col·laborat. Alguns socis permanents són la Biblioteca Nacional del País de Gal·les, JISC (Joint Information Systems Committee), *The Wellcome Library* o la Biblioteca de les Dones de la *Metropolitan University* de Londres.

L'Arxiu Web del Regne Unit duia a terme una estratègia selectiva de captura web, negociant amb les organitzacions i empreses propietàries abans de capturar la web, però durant el desenvolupament del present treball hi ha hagut un canvi legislatiu en matèria de dipòsit legal al Regne Unit que ha modificat l'escenari. Des del sis d'abril del 2013 una nova normativa amplia les capacitats de capturar i arxivar el web a sis biblioteques del país: la Biblioteca Britànica, la Biblioteca Nacional d'Escòcia, la Biblioteca Nacional del País de Gal·les, la Biblioteca Bodleiana, la Biblioteca de la Universitat de Cambridge i el *Trinity College* de Dublín.

Des de la Biblioteca Britànica han llançat tota una campanya a Internet i a les xarxes socials per demanar la participació de la societat britànica per definir els criteris de captura i crear un debat entorn a quines pàgines web s'han de preservar i quines utilitats se li poden donar. Assistim per tant a una nova etapa en la que l'Arxiu Web del Regne Unit veurà incrementat exponencialment el seu fons i veurem de quina manera els instruments institucionals de difusió aconseguen expandir dins la ciutadania el coneixement del *web archiving* i els beneficis que aquest pot tenir per la societat i el món empresarial britànic.

Existeixen molts més projectes d'arxivament web desenvolupats per biblioteques nacionals però no tenim temps de comentar-los tots, alguns són la Biblioteca Nacional de França, l'Arxiu web del Govern del Canadà, l'Arxiu web croata, l'Arxiu web islandès, l'Arxiu web d'Eslovènia, l'Arxiu web d'Àustria, etc.

6.2. Projectes supranacionals

Entenem per projectes supranacionals aquells que no se circumscriuen sota els interessos d'un país; projectes els interessos dels quals podríem considerar-los transnacionals que busquen assolir una missió que supera les fronteres.

El primer d'aquest projectes que hem de comentar, tant perquè va ser pioner en el camp de l'arxivament del web com perquè és el primer del món en quantitat d'informació capturada, és l'**Internet Archive**. És una fundació sense ànims de lucre que va ser creada l'any 1996 a San Francisco (EE.UU.) amb la finalitat de construir la Biblioteca d'Internet, que recopilés tot el coneixement i la cultura mitjançant la xarxa per evitar la seva pèrdua, i donar-li accés a les generacions futures i als investigadors, historiadors i estudiants del present.

L'*Internet Archive* recopila webs públiques de tot el món, de tots els dominis, en qualsevol idioma i proporciona accés públic a aquest contingut. La seva capacitat de captura és molt elevada i posseeix l'arxiu més gran del món que creix a un ritme de 100 terabytes al mes. Col·labora amb un gran nombre d'institucions que li encarreguen la captura i arxivament dels

seus propis webs, i d'altres, especialment universitats, fan servir el seu programari i el seu dipòsit per crear i gestionar les seves pròpies col·leccions.

Però l'*Internet Archive* no només arxiva pàgines web sinó tot allò que denomina artefacte cultural, i permet que qualsevol persona d'arreu del món pugui pujar contingut al seu repositori. Per tant, hi podem trobar pel·lícules, programes de televisió, programes de ràdio, e-books, etc. També treballa per incentivar la creació d'arxius web, i mitjançant la plataforma *Archive-it* ofereix serveis de captura a qualsevol que tingui interès en utilitzar-los. I organitza cada any el *K12 Web Archiving* que és una formació adreçada a joves estudiants per tal que aprenguin a crear les seves pròpies col·leccions de pàgines web.

Una altra organització de referència en matèria d'arxivament del web és l'***International Internet Preservation Consortium*** (IIPC), una institució internacional que treballa per la preservació del contingut d'Internet. Va ser fundada el juliol de 2003 quan comptava amb la participació de dotze membres, llista que s'ha anat ampliant considerablement fins a data d'avui, que són una trentena de socis aproximadament.

L'IIPC dona suport a les biblioteques nacionals per tal que engeguin projectes d'arxivament web, treballa per crear eines i tècniques comunes que possibilitin la creació d'arxius de pàgines web. Promou els estàndards entre els seus organismes membres per sumar esforços, i d'aquesta manera fer avançar els mètodes de captura desenvolupant recursos tecnològics cada cop més especialitzats. També és una plataforma que coordina diferents grups de treball que estudien aspectes com els mètodes de captura, les polítiques d'accés al contingut i els reptes de la preservació digital. L'IIPC publica periòdicament els resultats del seu treball i organitza conferències i reunions d'experts internacionals. També gestiona una llista de correu electrònic on es debaten totes aquestes qüestions.

Per últim comentarem el cas de l'***Internet Memory Foundation***, que és una fundació sense ànim de lucre amb seu a París i Àmsterdam, la missió de la qual és la preservació d'Internet. Col·labora amb múltiples organitzacions, i dona suport a la creació de col·leccions digitals que duen a terme institucions europees com la Biblioteca Nacional Central de Florència, el Parlament del Regne Unit o l'Organització Europea per a la Recerca Nuclear (CERN).

L'*Internet Memory Foundation* està involucrada en diversos projectes d'abast europeu, molts d'ells fundats per la Comissió Europea, que pretenen desenvolupar tecnologia i metodologia per la propera generació d'eines d'arxivament web. En definitiva, és un altre agent important en el desenvolupament i la difusió de l'arxivament web.

6.3. Universitats i organismes d'investigació

Altres institucions que treballen l'arxivament del web són els ens vinculats al camp de la recerca i la investigació. El món universitari a Estats Units és molt emprenedor i les universitats que disposen d'un elevat pressupost duen a terme empreses que les diferencien de la competència. Algunes han optat per projectes d'arxivament del web, que enriqueixen el seu patrimoni digital i les converteixen en institucions de referència a les que acudirán els investigadors donada la rellevància de les seves col·leccions. D'aquesta manera les universitats que lideren projectes d'arxivament web no només estan garantint l'accés a un contingut que d'una altra manera s'hauria perdut per sempre, sinó que s'estan convertint en subjectes actius que utilitzen el material capturat per crear un nou coneixement en benefici de la comunitat quan ofereixen accés als investigadors i experts perquè treballin i explotin les seves col·leccions digitals.

És el cas de la **Biblioteca de la Universitat de Columbia** que fa servir la plataforma *Archive-it* desenvolupada per l'*Internet Archive* per crear i gestionar les seves pròpies col·leccions. Les webs capturades pertanyen a alguns dels departaments de la Universitat o als seus afiliats, o bé són pàgines web d'individus o institucions que tenen documents o treballs custodiats a l'arxiu físic de la Universitat.

Un altre exemple en aquest sentit és el **Servei de Col·leccions d'Arxius Web** de la Universitat de Harvard, que va néixer el juliol del 2006 i va ser el primer projecte fundat per la *Library Digital Initiative* que estava orientat a l'estudi de la preservació a llarg termini de material nascut digital. Algunes de les col·leccions que gestionen tracten temes com la vida i l'estudi a Harvard, la història de les dones als Estats Units, o el debat creat al Japó entorn de la revisió de la Constitució vigent després de la Segona Guerra Mundial.

D'altra banda, la Universitat de Califòrnia mitjançant la seva **Biblioteca Digital de Califòrnia** ofereix serveis d'arxivament web a una llarga llista de col·laboradors, en la seva majoria relacionats amb el món universitari i de la recerca, perquè puguin crear, arxivar i gestionar les seves pròpies col·leccions digitals. D'aquesta manera els socis de la Biblioteca Digital de Califòrnia, alguns tant destacats com les Biblioteques de la Universitat de Nova York o la Biblioteca de la Universitat de Berkeley, poden capturar i mantenir sempre accessibles les pròpies pàgines web. I al mateix temps crear-ne col·leccions temàtiques d'aquelles matèries en les que siguin una referència en el món de l'educació i el coneixement.

6.4. Empreses

Dins de les empreses que fan servir l'arxivament web podem establir una clara divisió entre aquelles que es dediquen a oferir serveis d'arxivament web a tercers, i aquelles que fan servir la captura per arxivar les seves pròpies webs.

Des de l'origen de l'arxivament del web el 1996 i amb l'evolució tecnològica que va comportar han crescut tot un seguit d'empreses que fan del seu camp de negoci la captura, l'arxivament i l'emmagatzematge de pàgines web. Són empreses especialitzades en la gestió de sistemes d'informació i la comunicació per Internet que ofereixen eines i solucions de *web archiving* adaptables als interessos dels seus clients. Algunes són *Hanzo Archives*, *Nextpoint* o *Smarsh Web Archiving*.

D'altra banda, el percentatge d'empreses que realitza captures del propi web és mínim, però segur que mantindrà un creixement important durant els propers anys. Sobretot pel que fa empreses que tinguin en la seva web una eina important tant per comunicar al món els seus serveis, com aquelles que la facin servir com un instrument més de relació amb els seus clients. O també en la que es donin intercanvis d'informació rellevants que sigui important conservar, així com webs que ofereixen serveis d'oficina al estil seu electrònica.

7 L'arxivament del web a Catalunya: el PADICAT

A mitjan de la dècada passada la Biblioteca de Catalunya fa un replantejament dels mecanismes i les estratègies més adequades per preservar i difondre el patrimoni nacional en un món que ofereix noves possibilitats tecnològiques. Fruit d'aquesta reflexió l'any 2005 inicia quatre projectes innovadors:

- Col·laboració amb *Google Llibres* per la digitalització d'una part del fons bibliogràfic.
- Memòria Digital de Catalunya, és un repositori cooperatiu i d'accés obert, que conté col·leccions digitalitzades relacionades amb Catalunya i el seu patrimoni.
- Arxiu de Revistes Catalanes Antiques (ARCA): digitalització de revistes catalanes que ja no es publiquen.
- Patrimoni Digital de Catalunya (PADICAT), l'arxiu web de Catalunya.

El 2005 arrenca la fase de naixement del PADICAT, en la que es donen els passos preliminars d'estudi i preparació del projecte. El juliol del 2006 es comença a capturar el web de manera automatitzada, i el moment culminant arriba el setembre del mateix any amb l'estrena del portal web on es dona accés obert al material compilat.

Durant la fase de creixement (2007-2008) es treballa per executar els programes de desenvolupament traçats amb anterioritat i aconseguir un funcionament normal de totes les peces de l'engranatge: captura, indexació, arxivament, posada a disposició pública. En la fase de consolidació (2009-2011) es duu a terme una destacable tasca de difusió entre les institucions del país que dona com a resultat la signatura de convenis de col·laboració amb més de 450 organismes que cooperaran perquè el PADICAT capturi les seves pàgines web. L'11 de setembre de 2011 es modernitza el portal web d'accés, traduït a tres idiomes, i es milloren les opcions de cerca del catàleg possibilitant la consulta per matèries.

El PADICAT captura el patrimoni digital de Catalunya seguint el model híbrid de captura que hem comentat. Per una part captura massivament les webs d'accés públic del domini .cat, i per l'altra estableix acords amb institucions i organismes per procedir a la captura de les seves pàgines web. A més, realitza campanyes de captura puntuals vinculades a esdeveniments rellevants en la vida del país, com les campanyes electorals. A tall d'exemple, podem citar algunes de les monografies, que porten per títol: Eleccions europees (2009), Eleccions municipals (2007), Folk-rock (2008) o Museus de Catalunya.

El soci tecnològic indispensable en el projecte és el Centre de Serveis Científics i Acadèmics de Catalunya (CESCA), que ha desenvolupat i dissenyat algunes aplicacions per millorar l'accés i la recuperació dels recursos digitals dipositats al PADICAT. Aquests recursos es capturen mitjançant el programari *Heritrix*, desenvolupat per l'*Internet Archive* i les Biblioteques Nacionals Nòrdiques. I com a sistema de gestió documental per assignar metadades es fa servir una altra eina de codi obert, *Web Curator Tool*, creació conjunta de la Biblioteca Britànica i la Biblioteca Nacional de Nova Zelanda. Tots dos són programari lliure que gaudeixen d'un gran predicament entre la comunitat de biblioteques que capturen el web.

La consolidació d'aquesta infraestructura ha permès al PADICAT assolir uns nivells de captura que ells mateixos quantifiquen en la seva pàgina web institucional:

- Compilació semestral de 30.000 recursos del domini .cat.
- Compilació semestral de 550 recursos de les 450 entitats amb què s'ha arribat a un conveni de col·laboració.
- Compilació semestral dels 800 recursos procedents de recomanacions dels usuaris.
- Compilació única de 1.000 recursos de col·leccions monogràfiques.
- Compilació diària d'una part substancial de 30 publicacions seriadades en línia.

És una tasca molt valuosa que fins a dia d'avui ha aconseguit compilar un total de 13 terabytes d'espai, que reflecteixen una part important del patrimoni digital català.

El PADICAT és, per tant, el primer projecte d'arxivament web i de foment de la preservació digital que es dona a l'Estat Espanyol. No és fins el 2007 que es crea Ondarenet, l'arxiu electrònic del patrimoni digital basc, i fins el 2009 la Biblioteca Nacional de España no encarrega a l'*Internet Archive* les primeres captures del domini .es. En la seva curta existència, només set anys i mig, el PADICAT s'ha consolidat com un projecte capdavanter, que ha sabut construir-se una xarxa local de contactes que li garanteixi el recolzament institucional, i que també s'ha relacionat amb arxius digitals internacionals mitjançant la seva participació a l'*International Internet Preservation Consortium*, on ha pogut enriquir-se a partir de l'intercanvi d'experiències i coneixements de la resta d'institucions de *web archiving* del món.

Durant l'elaboració del present treball ens va semblar oportú poder comptar amb la opinió d'algun membre de l'equip del PADICAT, que ens oferís la seva visió experta sobre la pròpia institució i sobre algunes qüestions com la preservació digital, la llei de dipòsit legal o la col·laboració de les empreses i institucions amb el PADICAT. Ens vàrem adreçar al senyor Ciro Lluca, que és el coordinador del PADICAT, a més de professor associat de la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona. El senyor Lluca molt amablement es fa oferir a col·laborar i a contestar un breu qüestionari que tot seguit reproduïm.

Santi Lopera: Podria explicar breument quin és el procés de signatura d'un conveni de col·laboració entre una empresa o un particular i el PADICAT perquè aquest dugui a terme una captura periòdica del seu web?

Ciro Lluca: Un cop identificada l'empresa o institució amb la que volem establir un conveni, comuniquem el nostre sistema de funcionament, els beneficis del PADICAT, i trametem un model de conveni (<http://www.padicat.cat/ca/collabora-i-participa/estableix-un-conveni>). Si les negociacions prosperen, se signa per duplicat el conveni. A partir d'aquell moment passen uns tres mesos abans no es realitza la primera captura sencera, i amb posterioritat a aquella data la captura és semestral.

SL: Quines són les reaccions més freqüents per part de les institucions que són contactades pel PADICAT per capturar el seu web? Es dóna el cas que la empresa, o un particular, sigui qui faci el primer pas i es dirigeixi al PADICAT per demanar l'arxivament del seu web?

CL: Depèn de la naturalesa de l'empresa o institució i del seu coneixement sobre la Biblioteca de Catalunya. Alguns ajuntaments tanquen favorablement l'expedient en trenta dies (obrir l'expedient, passar el conveni per comissions informatives i plenari, trametre el conveni signat), altres triguen anys a fer les mateixes gestions, independentment de la mida o dels recursos que tinguin. El més ràpid de tots, per exemple, va ser Palafrugell (23.000 habitants). En tenim d'altres, molt més importants per població, que encara ho estan gestionant. En tot cas, les administracions públiques solen entendre el rol de la Biblioteca de Catalunya amb més facilitat que algunes empreses, que poden no concebre que el servei de la BC no té un cost directe per a elles (em refereixo, lògicament, a banda del que paguen amb els seus impostos). Recordo que Aigües de Barcelona, o Mango, per posar exemples

de grans empreses, varen trigar poquíssim a signar, un cop informades del procés. Altres, de la mateixa mida o més petites, han pogut trigar més, sovint perquè no hem estat capaços d'explicar molt sintèticament quin és l'objectiu del PADICAT.

SL: En línies generals, creu vostè que les empreses i institucions entenen la importància de la preservació digital o és necessària una forta tasca pedagògica per part del PADICAT?

CL: Com he llegit recentment, hi ha dos tipus de solucions a empreses: les vitamines (milloren el sistema de funcionament, a mig i llarg termini) i les aspirines (resolen un mal de cap existent). La preservació digital, tal com l'entendem majoritàriament, aporta beneficis a mig i llarg termini (vitamina). Sovint les empreses estan centrades en el dia a dia, i no poden —per recursos, per estratègia— pensar en qüestions futures com les que proposem (accés permanent a les seves pàgines web). Per tant, la pedagogia és imprescindible.

SL: Tenen dades de les visites i consultes al fons del PADICAT?

CL: 6.000 visites l'any 2012.

SL: Es té coneixement de l'existència de treballs d'investigació efectuats a partir del fons del PADICAT?

CL: Sí, essencialment de màrqueting i comunicació en ciència política, també d'escoles de disseny. I sobretot de temes de preservació digital, procedents dels estudis del Grau d'Informació i Documentació. També sovint ens han contactat estudiants del Màster d'Arxivística de la UAB per demanar alguna dada concreta o alguna qüestió relacionada amb la política del PADICAT.

SL: Tenen pensat desenvolupar (o ja s'estan aplicant) estratègies de difusió del fons del PADICAT i d'estimulació de la seva consulta per part de la comunitat investigadora i el públic en general?

CL: Treballem en dues línies. La primera és la difusió en entorns professionals d'institucions de la memòria (Biblioteques, Arxius, Museus), publicant articles en revistes científiques i professionals, presentacions en congressos, etc. També anem a universitats (ciència política) a explicar les col·leccions monogràfiques de les campanyes electorals, i impliquem els professors en la selecció de recursos.

La segona és difusió en el públic generalista, per mitjà de notes de premsa, presència a la Viquipèdia, les xarxes socials, etc.

SL: Ha afectat la crisi econòmica al desenvolupament del PADICAT? S'han buscat formes alternatives de finançament o acords de col·laboració amb altres institucions, com per exemple l'altre arxiu digital de la Península: l'ONDARENET?

CL: Des de l'inici del projecte es va implicar altres institucions: CESCO, per la part tècnica; i la Fundació puntCAT, per la difusió del projecte i la captura coordinada del .cat. A més, s'han establert diverses accions de col·laboració amb universitats (ESMUC per selecció de música folk-rock en línia, per exemple). El PADICAT, d'altra banda, és a la cartera de serveis de la Biblioteca de Catalunya per als projectes de patrocini que s'engeguen habitualment.

SL: L'*Internet Archive* mitjançant el seu servei *Archive-it* estableix acords amb diferents institucions (universitats, arxius, museus, etc.) perquè aquestes desenvolupin les seves pròpies col·leccions de pàgines web capturades, i fins i tot porta a terme plans de formació com el *K-12 Web Archiving* que pretén difondre la importància de la captura web. El futur de l'arxivament del web passa per conscienciar a la ciutadania que la responsabilitat de la preservació digital és una tasca compartida? Més enllà de les diferències entre el PADICAT i l'*Internet Archive*, es podrien dur a terme projectes similars al nostre país.

CL: Sí, és un bon model. Sens dubte el camí passa per la cooperació.

SL: Aquests dies estem veient com la Biblioteca Britànica està realitzant una gran campanya d'arxivament del web possibilitada per un canvi recent en la llei de dipòsit legal. Podria comentar breument si la nostra normativa afavoreix o dificulta la pràctica del web archiving?

CL: Amb sinceritat, penso que la pràctica del web archiving va molt més enllà de la regulació legal. De fet, en altres països (Països Baixos, Suïssa) ni existeix la llei del dipòsit legal, i el compromís amb la preservació per part de les institucions és molt elevada, en el sector analògic i en el digital. Vull dir, sense una política o estratègia, i sense recursos, la llei no serveix per a res.

Dit això, la llei espanyola és un bon text legal. La versió de 1971 era clarament desfasada, però la nova llei (23/2011, de 29 de juliol, que va entrar en vigor el gener de 2012) va per bon camí. Cal però que es desplegui via reial decret perquè sigui efectiva. En això hi estan treballant les diverses biblioteques nacionals que formen part del grup de treball de preservació digital del Consejo Español de Cooperación Bibliotecaria.

La llei espanyola beu, en qüestions de web archiving, d'un magnífic text legal: la llei danesa (<http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html>). La llei britànica, de fet, existeix des de 2003.

SL: Quins són els objectius i reptes del PADICAT pels pròxims anys?

CL: La missió del PADICAT és recollir, conservar i difondre el patrimoni digital de Catalunya nascut a Internet.

El sistema es basa en l'aplicació d'una sèrie de programes informàtics que permeten la captura, l'emmagatzematge, l'organització, la preservació i l'accés permanent a les pàgines web publicades a Internet.

Els seus objectius són:

- Compilar massivament el domini .cat.
- Impulsar el dipòsit sistemàtic de la producció web de les entitats i les empreses de Catalunya.
- Promoure línies de recerca processant de manera monogràfica els recursos d'esdeveniments de la vida pública catalana, com ara campanyes electorals a Internet, el fenomen de la música en línia, o els museus a Internet.

Després d'unes etapes de naixement (2005-2006), creixement (2007-2008) i consolidació (2009-2011), a partir del 2012 es persegueix sistematitzar la capacitat de creixement, amb la fita d'incorporar anualment unes 75.700 versions d'aproximadament 32.000 pàgines web, procedents de:

- Compilació semestral de 30.000 recursos del domini .cat.
- Compilació semestral de 550 recursos de les 450 entitats amb les que s'ha arribat a un conveni de cooperació.
- Compilació semestral dels 800 recursos procedents de recomanacions dels usuaris.
- Compilació única de 1.000 recursos de col·leccions monogràfiques.
- Compilació diària d'una part substancial de 30 publicacions seriades en línia.

A aquestes fites concretes s'hi afegeixen quatre eixos permanents de treball:

- Definició de les estratègies de preservació digital per al patrimoni nascut a Internet. PADICAT proporciona radiografies periòdiques de la web catalana; detecta els formats que experimenten a curt termini problemes d'il·legibilitat; identifica els llenguatges més usats, etc.
- Impuls a línies de recerca a partir de la creació de col·leccions monogràfiques que compten amb la implicació d'experts de cada matèria.
- Creació i manteniment de l'hemeroteca digital a Internet, amb la captura sistematitzada de publicacions digitals en sèrie. Actualment, una mostra representativa quant a tipus i continguts, seleccionant les nascudes digitals, sense equivalent analògic.
- Cooperació amb altres arxius web i dipòsits de preservació digital, de biblioteques, arxius i museus, per donar una resposta eficient als reptes de preservació digital i accés als recursos dipositats.

SL: Els arxivers treballem, custodiem i preservem documents d'arxiu, la definició oficial del qual és la següent: "És la informació enregistrada, amb independència de les seves característiques físiques i intel·lectuals, produïda o rebuda i conservada per una organització o una persona en el desenvolupament de les seves activitats". Creu que els arxivers hem de començar a pensar en les pàgines web com a documents d'arxiu? Creu possible un apropament a l'arxivament del web des de la metodologia arxivística?

CL: Em costa veure diferències entre els professionals que treballen a biblioteques i arxius i altres serveis d'informació, més enllà d'una procedència formativa que pot ser diversa i que, a la pràctica, no diferencia en pràcticament res la praxi professional. En molts països, arxivers i bibliotecaris treballen de manera conjunta per garantir la preservació del patrimoni,

sigui analògic o digital. A Catalunya ens ha costat més, però confio que les noves generacions d'arxivers i bibliotecaris se centrin en treballar pel servei públic i no per defenses gremials del segle passat. Cal sumar esforços, i a Catalunya, en aquest sentit, no hem fet tots els deures que havíem de fer.

PART II. EL *WEB ARCHIVING* I L'ARXIVÍSTICA

8 És el web un document d'arxiu?

Al principi del present treball advertíem que el terme “arxivament del web” havia estat donat pels bibliotecaris i professionals de la informació que, com hem vist, han sigut els principals protagonistes de l'origen i desenvolupament d'aquesta tècnica. Comentàvem que, tot i que agafa la mateixa paraula que dóna nom a la nostra disciplina, és una tècnica totalment aliena a la nostra professió. Però d'acord amb el que hem vist fins ara, realment és el *web archiving* una proposta estranya per a l'arxivística? Capturar, gestionar, conservar i fer accessibles evidències documentals no són tasques dels arxiviers?

La principal pregunta que ens hem de formular frontalment per saber si l'arxivament del web ha de ser una de les preocupacions dels arxiviers és la següent: són les pàgines web documents d'arxiu? Revisem la definició que la Norma de Descripció Arxivística de Catalunya dóna de document d'arxiu:

“És la informació enregistrada, **amb independència de les seves característiques físiques i intel·lectuals**, produïda o rebuda i conservada per una organització o una persona en el desenvolupament de les seves activitats”. (NODAC, pàg. 180)

Si atenem aquesta definició és evident que les pàgines web poden ser documents d'arxiu i, en conseqüència, els arxiviers hauríem d'intentar aproximar-nos al seu tractament des de la nostra pròpia metodologia. Una qüestió gens fàcil si tenim en compte el caràcter dinàmic de les pàgines web, que són objectes digitals complexos que canvien amb menor o major periodicitat. Però una qüestió, al cap i a la fi, que hem d'encarar des de la nostra experiència i coneixement teòric si volem ser protagonistes de la gestió documental del segle XXI.

Les pàgines web les han d'arxivar les biblioteques o els arxius? Fins ara són les biblioteques les que ho fan, pocs arxius treballen en aquest camp a nivell mundial. I no coneixem cap que ho faci al nostre país. Per mi, la resposta adequada a aquesta pregunta és que les pàgines web poden ser gestionades per biblioteques i també per arxius. Cada professional li donarà un tractament diferent, perquè el seu objecte d'estudi i la seva metodologia són diferents. Les biblioteques busquen preservar la producció digital del país, atresorar les expressions culturals i científiques de la societat perquè les generacions futures les puguin estudiar i conèixer.

Però els arxius han de preocupar-se de les evidències documentals que les pàgines web contenen. Els arxius no han de dirigir els seus ulls a Internet com si fos un immens oceà al que han de llençar les seves xarxes per veure què capturen. No és la manera com treballen els arxius. Els arxius ho són en funció de l'organització per la que treballen, no tracten col·leccions, sinó fons documentals que, com sabem, vénen determinats per un productor. Te aquest productor una web corporativa? La fa servir com un aparador per comunicar al món la seva imatge? O la fa servir com un instrument a través del qual relacionar-se amb els seus clients i proveïdors? Es dóna intercanvi d'informació en aquesta pàgina web? És aquesta informació una evidència de les activitats desenvolupades per l'organització?

És aquest el procés que considero que han de fer els arxius amb el *web archiving*. Han d'agafar aquesta tècnica, descompondre-la i tornar-la a muntar amb criteris arxivístics. Com hem vist, és una disciplina que durant disset anys ha evolucionat sota uns criteris determinats dintre dels quals no hi eren els arxivístics, però que te molt per oferir a la professió. Les seves eines seran molt útils quan, tard o d'hora, els arxivers haguem de tractar la qüestió de les pàgines web perquè l'evolució de la comunicació passa per Internet de forma ineludible. Per tant, donat que és un camí que hem de recórrer, ja fem tard en no començar-lo a caminar.

9 Per què els arxius de Catalunya no arxivem els webs?

Una anàlisi profunda sobre la història de l'arxivística a Catalunya supera l'abast d'aquest treball, i els coneixements del seu autor, però sí voldria mirar d'apuntar algunes possibles explicacions que responguin a la pregunta de per què els arxiviers no arxivem les webs. Sobretot és un intent de respondre'm a mi mateix aquest interrogant, que fins ara no ha estat plantejat, tenint en compte que els criteris científics i professionals ens emparen per poder desenvolupar aquesta tasca amb el convenciment que forma part del nostre camp d'actuació, i que és responsabilitat de l'arxiver i del gestor documental tractar arxivísticament les evidències documentals que contenen les pàgines web.

No hi ha una única resposta que contesti aquest interrogant, sinó que serà la combinació de tot un seguit de factors el que podria explicar aquesta situació. Per un costat podríem dir que Catalunya és un país amb una gran tradició arxivística però que no ha sigut fins una data molt recent que no ha vist professionalitzada la seva disciplina. S'ha fet molta feina des de l'arribada de la democràcia per desenvolupar un sistema arxivístic cohesionat que arribés a tot el territori, aprovant normatives i lleis per garantir la integritat del patrimoni documental. S'han establert pautes per actuar amb la documentació, s'han marcat criteris de restauració i conservació, s'han definit perfils professionals, s'ha implantat un sistema de gestió documental que contempla arxius de gestió, arxius intermedis i arxius històrics. S'han creat institucions cabdals com l'Arxiu Nacional de Catalunya, que vetlla per la conservació del patrimoni documental català, o la Comissió Nacional d'Accés Avaluació i Tria Documental, que ha de definir el període de retenció i el règim general d'accés dels documents. S'ha fet molta i bona feina en un període de temps relativament curt, i aquesta feina s'ha de reconèixer, però l'esperit crític serà el que ens farà avançar i millorar. I existeixen símptomes que ens indiquen que la professió arxivística a Catalunya encara no està plenament normalitzada: no existeix un col·legi professional d'arxiviers (sí una Associació d'Arxiviers i Gestors Documentals), no existeix una carrera universitària d'arxivística homologable fora del país (sí l'Escola Superior d'Arxivística i Gestió Documental), la consideració que l'Administració té de l'arxivística és menor i així es demostra quan l'arracona en l'àmbit cultural en comptes de reconèixer-li el seu caràcter transversal, que veritablement li permetria incidir en la gestió documental de les institucions públiques.

No perdo de vista el tema del present treball quan parlo de tot això. El caràcter lateral que l'Administració atorga a l'arxivística és probablement la raó per la qual fa bastants anys que hi ha una assignatura pendent en aquest país com és l'administració electrònica. I de retruc, és clar, l'arxivament del web que, com hem vist, no és imaginable sense un sistema de

gestió de documents electrònics previ. Els gestors documentals tenim molt poca incidència en les solucions tecnològiques que fins ara s'han buscat a l'Administració i molt freqüentment es vol construir la casa per la teulada evitant els passos intermedis. És sabut que per implantar un sistema de gestió de documents electrònics s'han d'automatitzar els processos de treball. Per automatitzar-los, prèviament s'han de diagramar per entendre els fluxes de feina i la producció documental que generen. I abans encara, i aquí està la clau de volta de tot al nostre parer, per poder diagramar els processos és necessari entendre com treballa una organització fins en el més ínfim detall. És una tasca dura de formigueta, que requereix de molt temps i que implicar dialogar amb els treballadors implicats en cada procediment per copsar bé de quina manera es funciona. Aquesta no és una tasca que generi resultats a curt i mig termini, i és molt costosa en temps i esforços. Quantes organitzacions, corporacions, agències i entitats públiques (o empreses privades) coneixem que tinguin diagramats i automatitzats tots els seus procediments?

Qualsevol solució tecnològica que es busqui requerirà d'una adaptació a la manera de treballar de la institució, i aquest coneixement previ l'hem d'aportar els arxivistes i gestors documentals, que coneixem com ningú altre les funcions i activitats de l'organisme pel qual treballem. Considero que fins que aquest pas no s'hagi fet l'administració electrònica no acabarà de funcionar, de la mateixa manera que un edifici no pot aguantar-se sense uns fonaments sòlids. I fins que l'Administració no s'adoni, o els gestors documentals la fem adonar-se, que el camí correcte passa per dotar els serveis d'arxiu de més capacitat d'acció i protagonisme en el funcionament de les entitats no aprovarem aquesta assignatura.

Altres raons que expliquen el fet que els arxius actualment no arxivin les pàgines web també les podem anar a buscar dintre de la mateixa professió. Potser per aquesta adscripció equivocada de l'arxivística a l'àmbit cultural, o pel tant per cent elevat d'arxivistes que provenen de la branca d'estudis d'Humanitats (jo mateix sóc llicenciat en Història) hi ha, dins de la professió, certa resistència a les noves tecnologies. He tingut l'ocasió de comprovar-ho entre els companys de l'ESAGED i, recentment, ha tornat a sortir aquest tema de discussió a la llista de distribució Arxiforum¹⁴. Es podria fer de forma esquemàtica una divisió entre una tendència més historicista, que valora la importància cultural i històrica dels documents i entén l'arxiu com un edifici en el qual custodiar els documents de gran significació cultural, i una altra que defensa la importància de la gestió documental com un conjunt de principis i metodologies necessaris pel bon funcionament d'una organització.

Segons el meu parer, qui pensi que la figura de l'arxiver ha de tenir poca relació amb la tecnologia s'equivoca de bat a bat. Amb tota seguretat els primers arxivistes de la Història

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'Arxivistes – Gestors de Documents de Catalunya.

eren escrivans, i els escrivans eren experts en la tecnologia de l'escriptura, que no era quelcom que estigués a l'abast de tothom. Coneixien perfectament el procés de preparació de les pells per fer els pergamins, coneixien la composició de les tintes que feien servir i la varietat que existia, i el temps que necessitaven per assecar-se, així com els estris més adients per escriure. Per tant, un requisit imprescindible de la professió d'arxiver i gestor documental hauria de ser el de conèixer la tecnologia que genera els documents que ell custodia. Naturalment que l'esclat de les noves tecnologies de la informació i la comunicació no ho posa fàcil, però en comptes de resistència com a mínim hi hauria d'haver un interès en aprendre el funcionament d'aquests nous mitjans. Falta obertura de mires en aquest sentit, i falta formació específica que ajudi els arxivers i gestors documentals a fer front a reptes com el de la preservació digital o el *web archiving*.

10 Incorporació de l'arxivament web a les tasques d'arxiu

Abordar el repte de desenvolupar tasques d'arxivament web des dels arxius no és quelcom diferent a encarar la gestió de documents electrònics. Pràcticament és el mateix, es podria dir que només és anar un pas més enllà i significaria atorgar a les pàgines web la importància que aquestes ja han adquirit en la nostra societat. Les pàgines web haurien de ser tractades com si fossin documents electrònics —i, de fet, ho són— en la gestió dels quals el tractament de les metadades és un factor absolutament clau.

Un primer pas ineludible seria arribar a un consens sobre quines parts del web es consideren documents d'arxiu; si el web sencer també el podem considerar com a document tot ell, o només els objectes digitals que contingui. És necessari fer aquesta reflexió teòrica per establir quines seran les unitats bàsiques amb les que treballarem posteriorment. De ben segur que les propostes seran moltes i molt variades, si atenem a la complexitat que les pàgines web han anat adquirint amb l'evolució tecnològica. Han passat de ser, en un primer moment, simples imatges fixes fins esdevenir complexes bases de dades interconnectades. Serà qüestió, potser, que cada institució defineixi, en funció de l'ús que li dóna a la seva pàgina web, quines parts d'aquesta constitueixen evidències de les seves activitats.

Aquest és un debat que els Arxius Nacionals d'Austràlia ja van tenir fa més de 10 anys i com a resultat van publicar unes recomanacions per gestionar documents d'arxiu basats en el web¹⁵. Aquestes recomanacions contempen dues estratègies diferents alhora de tractar els documents basats en el web:

-*Object-driven approach*. Que podríem traduir com «aproximació basada en l'objecte». Més adient per a webs estàtiques, que funcionen com a magatzems d'objectes digitals. Aquestes pàgines web podrien capturar-se en la seva totalitat conservant una panoràmica global (*snapshot*) i, amb posterioritat, fer un seguiment automatitzat dels canvis que es donin en el temps: nous objectes afegits, objectes modificats, etc.

-*Event-driven approach*. Aproximació basada en l'esdeveniment. És un tractament propi per pàgines dissenyades per interaccionar amb l'usuari. Aquestes pàgines web acostumen a treballar amb bases de dades de perfils d'usuari, mecanismes de cerca, accions programades i altres funcionalitats. Per tant, seria necessari poder enregistrar totes les

¹⁵ National Archives of Australia

accions que un usuari efectua dins de l'entorn web, capturar la seva adreça IP o domini, el perfil d'usuari, hora i data de les accions, etc.

Un cop estigués definida aquesta qüestió, un altre pas a considerar seria el moment en el qual aquests documents basats en el web han de ser capturats i introduïts en el sistema de gestió documental de l'organització. La resposta hauria de donar-se en relació a la funció d'aquests documents. Si van lligats a un expedient o a un tràmit dins del qual constitueixen una evidència o tenen una importància com a peces informatives, les pautes de quan han de ser capturades vindrien donades per la vida del tràmit. Per una altra part, si les captures només són panoràmiques, és a dir, es realitzen per testimoniar l'evolució de la pàgina web, es podrien automatitzar a intervals de temps regulars. En funció del nivell d'activitat del web podrien ser setmanals, mensuals, trimestrals, etc.

Una vegada aquests documents provinents del web han entrat en el sistema de gestió documental, el seu tractament seria el mateix que el de la resta de documents electrònics. Serien classificats atenent al quadre de classificació de la institució, i se'ls aplicarien les polítiques de disposició, de seguretat i d'accés pertinents.

El requisit previ, per tant, perquè una organització pugui desenvolupar tasques d'arxivament de les seves pàgines web és que disposi d'un sistema de gestió documental orientat al tractament de documents electrònics, amb totes les eines que això requereix: quadre de classificació, calendari de disposició, esquemes de metadades, quadre de seguretat i accés, etc.

11 Escenaris possibles

Si fem l'exercici d'imaginar de quina manera els arxius públics de Catalunya poden començar a integrar tasques de *web archiving* en el seu treball diari se'ns acudeix que podrien donar-se dos possibles escenaris de futur.

Xarxa descentralitzada d'arxius web

L'escenari més probable podria ser aquell en el qual els arxius amb més capacitat (pressupostària, tecnològica, etc.) engegessin els seus propis projectes d'arxivament web. Serien arxius que ja disposen d'un sistema de gestió de documents electrònics implementat, i que voldrien anar un pas més enllà i començar a capturar, tractar i custodiar les pàgines web de la seva institució.

Seria desitjable que aquestes institucions publicuessin un reglament propi que defineixi quins són els documents web que han de ser capturats i preservats per garantir els interessos de la pròpia organització. Incorporar aquestes noves responsabilitats no hauria de suposar un gran impediment per a un servei d'arxiu que ja treballi amb documents electrònics. Es necessitaria un programari de captura i gestió de pàgines web, el coneixement necessari per fer-lo servir i un emmagatzematge digital suficient. Però per a una institució que ja compti amb un repositori digital segur ampliar la capacitat d'emmagatzematge no ha de ser un problema, sobretot perquè aquest no ha de ser molt elevat donat que la porció d'Internet que es vol capturar és mínima. Són les biblioteques nacionals les que volen conservar grans quantitats d'informació; els arxius en tindrien suficient amb els documents de les pàgines web de la seva institució.

Aquest model descentralitzat faria més gran la bretxa digital que actualment ja existeix en el Sistema d'Arxius de Catalunya, on es donen situacions disperses d'arxius que només treballen en paper i altres que s'han pogut dotar d'eines més avançades que els permeten gestionar documents electrònics.

Arxiu web central, sota el paraigües institucional de l'ANC

L'altre escenari possible seria la creació d'un repositori institucional centralitzat que donés servei a tot el Sistema d'Arxius de Catalunya. Igualment seria imprescindible l'aprovació i publicació d'un reglament que marqués els principis i les pautes d'actuació de la captura web. O millor que un reglament seria una normativa fruit del consens de comissions d'experts i grups de treball, amb voluntat globalitzadora.

Cada arxiu del SAC capturaria els seus propis recursos web però aquests serien emmagatzemats en un repositori centralitzat compartit amb tota la xarxa d'arxius. Un gestor documental implementaria les metadades necessàries que identificarien l'organització generadora de les captures i facilitaria la posterior recuperació de la informació. Això obligaria a iniciar un procés d'equiparació tecnològica d'aquells arxius que fins aleshores no han tingut l'oportunitat de dotar-se d'aquests mitjans. Caldria realitzar la instal·lació del programari necessari per la captura en cada arxiu, i garantir els recursos tecnològics necessaris perquè les captures arribin per un canal segur al repositori central. Tecnològicament això és relativament senzill en l'actualitat i no requeriria d'un maquinari especial, ja que existeixen mecanismes per crear canals segurs d'intercanvi d'informació per Internet fent servir sistemes d'encryptació específics.

Per tant, la inversió econòmica seria continguda tenint en compte que existeix un programari lliure de captura web, i que els arxius només necessitarien un ordinador amb connexió a Internet. La part forta de la despesa aniria destinada a la creació d'un gran repositori central, amb tots els requisits tecnològics necessaris, maquinari i programari específic, seguretat de les instal·lacions, etc.

En aquest model es garantiria una major cohesió entre els membres del Sistema d'Arxius de Catalunya, establint uns mínims protocols d'actuació i estandarditzant unes pràctiques comunes emanades d'un òrgan central director, que coordinés els esforços i impulsés les polítiques definidores del futur de la pràctica arxivística del país.

12 Conclusions

Al llarg de la història de la humanitat es donen moments en els quals les innovacions tecnològiques impacten de tal manera sobre el comportament de les societats que es fa difícil copsar la dimensió de la seva importància. Freqüentment es requereix de perspectiva històrica per poder avaluar de forma més encertada no només de quina manera l'avenç tecnològic va revolucionar la vida material de l'ésser humà sinó, i el que és més important, de quina manera li fa ampliar el seu horitzó de pensament. Com polvoritza els esquemes mentals i permet a les comunitats humanes pujar un graó més en l'escala evolutiva. Exemples d'això són el descobriment del foc en la nit des temps, la invenció de la roda, el desenvolupament de l'astrolabi, la creació del telescopi, la invenció de la impremta, la fabricació de la màquina de vapor, etc.

Totes aquestes tecnologies van produir canvis tan incomparables que les societats humanes van requerir d'un temps per assimilar-les i poder adaptar-se al nou món de possibilitats que obrien. De la mateixa manera, la nostra societat del segle XXI està adaptant-se a la irrupció de les noves tecnologies de la informació i la comunicació i, especialment, al naixement d'Internet que ho està canviant tot. I encara no podem copsar de quina manera tan profunda està movent les arrels del món contemporani. Les possibilitats de la comunicació instantània amb qualsevol part del món, les enormes quantitats d'informació disponibles per qualsevol persona amb accés a la xarxa, l'enorme poder organitzatiu de les xarxes socials, etc.

No caurem pas en la ingenuïtat de creure que Internet és la solució a tots els problemes de la humanitat; és un mitjà, i l'ésser humà és el responsable de l'ús que en faci d'ell. Eugeny Morozov ja ha advertit de la important eina de control que per als règims autoritaris suposa Internet i del perill del *ciber-utopisme* que creu que universalitzant l'accés a Internet es resoldran les desigualtats socials. Tot això forma part del període d'adaptació al nou paradigma tecnològic, estudiar les possibilitats que ofereix, i entendre els perills que pot suposar.

Les societats que més ràpidament i millor s'adaptin als canvis tecnològics tindran molt guanyat. Durant el present treball hem vist com una professió com la de bibliotecari ha reaccionat a la irrupció d'Internet i ha aconseguit entendre les seves responsabilitats dins d'aquesta nova situació. Ha reorientat les seves prioritats i ha desenvolupat l'arxivament del web, que és la tecnologia que els permet continuar desenvolupant les seves obligacions de custodis de la cultura i del patrimoni editorial nacional en el nou paradigma tecnològic.

Els arxivers hem d'entendre que els sistemes de producció de documents han canviat i, sense deixar de banda el patrimoni documental analògic que s'ha de continuar tractant, custodiant i explotant, hem d'adaptar-nos a la gestió del document electrònic. Podem trobar

elements d'utilitat en el *web archiving* que ens ajudin a fer aquest procés d'adaptació, donat que cada cop més els documents també es produeixen en les pàgines web. Estudiar i analitzar la tecnologia i els procediments que els bibliotecaris i enginyers de la informació han desenvolupat per capturar les pàgines web ens pot ajudar a construir la nostra pròpia manera de capturar els documents que es creen i s'intercanvien en les pàgines web. Sota la nostra pròpia metodologia i els nostres principis teòrics, per mantenir aquestes evidències documentals íntegres, autèntiques, recuperables i utilitzables. No existeixen raonaments científics que impedeixin als arxivers tractar les pàgines web com a documents d'arxiu.

Aquest pas que hem de donar els arxivers com a professió és una fase més de l'adaptació a la gestió de documents electrònics. Per poder anar a buscar documents a les pàgines web per integrar-los en el sistema de gestió documental de les nostres organitzacions, abans, encara que resulti evident dir-ho, hem d'haver implantat plenament un sistema propi de gestió de documents electrònics. I aquest és el punt on encara ens trobem avui dia. Tenim el coneixement, i sabem quines eines necessitem. Per tractar la documentació analògica en teníem prou amb algunes eines com el quadre de classificació i alguns instruments de descripció que ens ajudessin a recuperar els documents. Però per assumir plenament la gestió de documents electrònics hem de diagramar els processos i automatitzar-los, és a dir, reflexionar profundament al voltant de com i per què treballem.

I fins que aquest pas no s'hagi donat a totes les peces del Sistema d'Arxius de Catalunya no serà possible la creació d'un arxiu web centralitzat sota el paraigües institucional de l'Arxiu Nacional de Catalunya, com apuntàvem en el punt 4 de la segona part del treball. Perquè no és una qüestió de pura inversió tecnològica, si no de reconfigurar les maneres de treballar de cada institució. I els arxivers i gestors documentals hauríem de tenir la responsabilitat de liderar i impulsar aquests canvis organitzatius. Per això és imprescindible que, per un costat, els arxivers i gestors documentals millorin les seves competències tecnològiques per entendre més profundament els nous sistemes de producció documental digital. I, per un altre, poder gaudir d'un millor reconeixement per les cúpules de les organitzacions que els atorgui més poders i capacitats decisòries per propiciar aquests canvis en el si de les organitzacions des del coneixement expert que només un arxiver i gestor documental pot tenir.

13 Glossari

Open-source: terme que s'ha traduït al català com codi obert. Fa referència al programari distribuït i desenvolupat de forma lliure. La seva característica és que dóna accés al codi en el qual ha estat programat i això pot permetre que els usuaris l'adaptin a les seves necessitats i col·laborin en el seu desenvolupament.

Backup: és una còpia de seguretat que es fa d'unes dades amb la finalitat de poder recuperar-la en cas de pèrdua de la informació original.

Ciber-utopisme: és un terme que va crear el pensador i escriptor Eugeny Morozov en la seva obra *The Net Delusion: The Dark Side of Internet Freedom* (*La il·lusió de Internet: el lado oscuro de la libertad de Internet*) que fa referència a la creença que les comunicacions per la xarxa i Internet afavoreixen els oprimits més que els opressors.

Cookies: és el terme que designa una petita part d'informació que envia un lloc web i emmagatzema el navegador de l'usuari, de manera que el lloc web pot consultar l'activitat prèvia de l'usuari.

URL: de l'anglès *Uniform Recurs Locator*. És una cadena de caràcters que informa al navegador de la màquina on és el recurs al qual fa referència; el protocol que s'ha d'utilitzar per obtenir aquest recurs; i la manera com el servidor web trobarà quin és el recurs.

Web crawler: (també conegut com aranya web) és un programari que inspecciona les pàgines web de forma automatitzada. La seva funció acostuma a ser la de realitzar una còpia de les pàgines web, que posteriorment seran indexades per facilitar la seva recuperació.

14 Bibliografia i fonts

Ayre, Catherine; Muir, Adrienne. "The right to preserve: the rights issues of digital preservation". *D-lib magazine*. Volum 10, núm. 3. (Març de 2004). En línia <<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 línia <<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 línia <<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 línia <<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*. (6é, 2008, Alice Springs). En línia <<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 initiatives". *International Conference of Theory and Practice of Digital Libraries* (15enes, 2011, Berlin). En línia <<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, febrer 1997, París. En línia <<http://www.ica.org/11878/digital-recordkeeping-programme-resources/.html>>

Kahle, Brewster. "Editors' Interview: The Internet Archive." *RLG DigiNew*. Volum 6, núm. 3. (15 juny 2002). En línia: <<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* (63enes, 1997, Copenhagen). En línia: <<http://archive.ifla.org/IV/ifla63/63kuny1.pdf>>

Llueca, Ciro. "Webs sempre accessibles: las bibliotecas nacionales y los depósitos digitales nacionales". *BiD: textos universitarios de biblioteconomía i documentación*. Núm. 15 (deseembre de 2005). En línia <<http://www.ub.edu/bid/15lluca1.htm>>

Llueca, Ciro. "Archivando la Web, el proyecto Padicat (Patrimonio Digital de Cataluña)". *El profesional de la información*. Volum 15, núm. 6 (2006), p. 473-478. En línia <<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* (11es, 22-23 maig del 2008, Barcelona). En línia <<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 línia <<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 línia <<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*, (12enes, 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*. Volum 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". Presentació al curs 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* Volum 8, núm. 12 (deseembre 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*, (4tes, 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* (15enes, 2006, Edinburg). En línia: <<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* (10nes, 2008, Madrid). En línia <<http://eprints.rclis.org/12553/>>

Pulgar Vernalte, Francisca; Marcos Maciá, Sonia. "Capturing the Basque Web". *Libraries in the Digital Age LIDA*, (2nes, 2009, Dubrovnik i Zadar). En línia <<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 d'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 línia:<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 línia <<http://bit.ly/U2whWh>>

15 Annexos

Llistat de membres del International Internet Preservation Consortium. No constitueixen totes les iniciatives d'arxivament web existents però sí són de les més destacades.

Institució - Nom de l'arxiu		Categoria
Biblioteca Alexandrina		Fundació sense ànim de lucre
URL	http://www.bibalex.org/isis/frontend/archive/archive_web.aspx	
Biblioteca Nacional de França		Biblioteca Nacional
URL	http://www.bnf.fr/fr/collections_et_services/collections_departements.html	
Biblioteques de la Universitat de Columbia		Universitat/Recerca
URL	http://www.archive-it.org/organizations/304	
Biblioteca i Arxius de Canadà - Arxiu Web del Govern de Canadà		Biblioteca Nacional
URL	http://www.collectionscanada.gc.ca/	
Biblioteca de Harvard - Servei d'arxiu web de col·leccions de la Universitat de Harvard		Universitat/Recerca
URL	http://wax.lib.harvard.edu/collections/home.do	
Biblioteca Nacional i Universitària de Croàcia - Arxiu Web Croata		Biblioteca Nacional
URL	http://haw.nsk.hr/	
Institut Nacional de l'Audiovisual (França)		Biblioteca Nacional
URL	http://www.ina.fr/	
Internet Archive		Fundació sense ànim de lucre
URL	http://archive.org/index.php	
Internet Memory Foundation		Fundació sense ànim de lucre
URL	http://internetmemory.org/en/	

Biblioteca Nacional i Universitària d'Islàndia - Arxiu Web Islandès	Biblioteca Nacional
URL	http://vefsafn.is/
Biblioteca Nacional de Finlàndia - Arxiu Web Finlandès	Biblioteca Nacional
URL	http://webarchive.nationallibrary.fi/
Biblioteca Nacional de Suècia - Kulturarw3	Biblioteca Nacional
URL	http://www.kb.se/om/projekt/Svenska-webbsidor---Kulturarw3/
Biblioteca del Congrés dels Estats Units - Arxiu Web de la Biblioteca del Congrés	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 - Arxiu Web de Noruega	Biblioteca Nacional
URL	http://www.nb.no/fag/nasionalbibliotekets-samling/nettdokumenter2
Biblioteca Nacional de Nova Zelanda - Arxiu Web de Nova 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 d'Austràlia - PANDORA	Biblioteca Nacional
URL	http://pandora.nla.gov.au/
Biblioteca de Catalunya - PADICAT	Biblioteca Nacional
URL	http://www.padicat.cat/
Biblioteca Nacional d'Espanya - Arxiu Web Espanyol	Biblioteca Nacional

URL	http://www.bne.es/es/Inicio/index.html	
	Biblioteca Nacional i Universitària d'Eslovènia - Arxiu Web d'Eslovènia	Biblioteca Nacional
URL	http://www.nuk.uni-lj.si/	
	Els Arxius Nacionals del Regne Unit - Arxiu Web del Govern del Regne Unit	Biblioteca Nacional
URL	http://www.nationalarchives.gov.uk/webarchive/	
	Biblioteca Britànica - Arxiu Web del Regne Unit	Biblioteca Nacional
URL	http://www.webarchive.org.uk/ukwa/	
	Biblioteca Nacional Diet - Projecte d'Arxiu Web del Japó	Biblioteca Nacional
URL	http://warp.da.ndl.go.jp/search/	
	Biblioteca Digital de Califòrnia - Servei d'Arxiu Web	Universitat/Recerca
URL	http://webarchives.cdlib.org/	
	Biblioteca Nacional dels Països Baixos - Arxiu Web dels Països Baixos	Biblioteca Nacional
URL	http://www.kb.nl/	
	Biblioteca Nacional de la República Txeca - Arxiu del Web Txec	Biblioteca Nacional
URL	http://webarchiv.cz/	
	Biblioteca Nacional d'Àustria - Arxiu Web d'Àustria	Biblioteca Nacional
URL	http://www.onb.ac.at/ev/about/webarchive.htm	
	Biblioteca Nacional de Suïssa - Arxiu Web de Suïssa	Biblioteca Nacional
URL	https://www.e-helvetica.nb.admin.ch/pages/main.jsf	