El PADICAT, l'experiència catalana en l'arxiu d'Internet

CIRO LLUECA, DANIEL CÓCERA. Biblioteca de Catalunya
NATALIA TORRES, GERARD SUADES, RICARD DE LA VEGA.
Centre de Supercomputació de Catalunya

L'arxiu d'Internet

L'ús d'Internet es va generalitzar en els països desenvolupats a partir de mitjan anys noranta. Des d'aleshores, les tecnologies de la informació i la comunicació han facilitat que el patrimoni cultural i científic i la resta d'informació es presentin en format digital i, en conseqüència, es produeixi un creixement exponencial dels recursos digitals publicats en línia. Tal com ho exposava la UNESCO en les seves directrius per a la preservació del patrimoni digital, els recursos que són fruit del coneixement o l'expressió dels éssers humans, de caràcter cultural, educatiu, científic o administratiu, o que comprenen informació tècnica, jurídica, mèdica i d'altres tipus, es generen cada vegada més sovint directament en format digital, o es converteixen a aquest format a partir de material analògic ja existent.[1]

En paral·lel a l'aparició progressiva de servidors i pàgines web a Internet, les administracions públiques de diversos països han dissenyat estratègies per

garantir l'accés als continguts publicats en línia i la seva preservació per mitjà de la captura i el processament corresponents.

El repte no és menor. A més de la inexistència generalitzada d'un text legal actualitzat que doni cobertura legal a aquests processos documentals, actualment no hi ha sistemes informàtics que executin impecablement les operacions de compilació, processament i difusió dels recursos digitals en un entorn, Internet, que és dinàmic per definició. Malgrat aquestes dificultats, diversos països estan portant a terme accions sistemàtiques de preservació de la producció digital més òbvia: les pàgines web, per mitjà de la creació de dipòsits digitals anomenats comunament *arxius web*.

El benefici que generen aquests repositoris contemporanis és inherent a l'acció de les institucions de la memòria —biblioteques, arxius i museus—, la garantia de l'accés permanent al patrimoni creat per una comunitat, per contribuir al progrés i el creixement individual i col·lectiu dels seus membres.

Tendències

Hi ha nombrosos dipòsits digitals destinats a l'arxiu d'Internet en funcionament, a més d'una extensa bibliografia que els ha detallat i analitzat.[2] Els més coneguts són també els que van fer les primeres passes l'any 1996: el suec Kulturarw3 i l'australià Pandora, i un conegut repositori d'abast internacional, l'Internet Archive. Quinze anys més tard, podem comptar fins a cinquanta projectes en diverses fases d'implementació, tot i que només un terç d'aquesta xifra són accions consolidades.[3]

L'anàlisi d'aquestes experiències mostra dos models bàsics de polítiques de col·lecció amb una tendència generalitzada cap a un model híbrid. El primer és el model integral o exhaustiu (majoritari, i característic dels països escandinaus en els inicis), que persegueix la integració automàtica de la web a partir de determinats criteris infraestructurals (segons el domini de les pàgines web, segons la ubicació del servidor, etc.). El segon model és el selectiu (assimilat per Austràlia, el Regne Unit o el Japó, entre altres comunitats), dirigit a compilar la web basant-se en una política selectiva (un repertori de recursos digitals corresponents a les diverses àrees del coneixement per a un espai geogràfic concret). Aquests dos models clàssics han donat pas —en el que és a partir de l'experiència inicial danesa, una tendència generalitzada arreu del món— a models híbrids, que complementen la captura periòdica d'un domini geogràfic sencer, amb accions selectives, i amplien aquesta cobertura a diversos esdeve-

niments d'interès social (eleccions, competicions esportives, guardons culturals) o successos informatius que generen activitat intensa a les xarxes (atemptats, catàstrofes naturals, pandèmies, episodis de la crisi econòmica, debats socials).

Lamentablement, el nombre de dipòsits que permet accedir lliurement a les seves col·leccions és molt limitat, bé per evitar conflictes amb la vulneració dels drets de propietat intel·lectual dels recursos capturats sense autorització expressa, bé perquè les interfícies de recuperació de la informació dipositada no han estat prou desenvolupades.

En la major part dels casos han estat impulsors d'aquests projectes els organismes nacionals de biblioteques i arxius, a més de diverses entitats públiques i privades d'abast nacional o internacional. Representants d'aquests organismes procedents d'Alemanya, Austràlia, Àustria, el Canadà, Catalunya, Corea, Croàcia, Dinamarca, Escòcia, Eslovènia, Espanya, els Estats Units, Finlàndia, França, Israel, Islàndia, Itàlia, el Japó, Noruega, Nova Zelanda, els Països Baixos, Polònia, el Quebec, el Regne Unit, Singapur, Suècia, Suïssa i Txèquia s'agrupen en l'International Internet Preservation Consortium (Consorci Internacional per la preservació d'Internet, IIPC per la sigla anglesa) amb la missió de compilar i preservar la informació i el coneixement d'Internet, donar-hi accessibilitat, per a futures generacions de tot el món, i promoure l'intercanvi global i les relacions internacionals.

A Espanya, la Biblioteca de Catalunya (BC) va iniciar el 2005 el projecte PADICAT (Patrimoni Digital de Catalunya),[4] dedicat a l'arxiu sistemàtic de la Internet catalana.[5] El 2007, el Govern basc i Eusko Jaurlaritzaren Informatika Elkartea (EJIE, Societat Informàtica del Govern Basc) van crear Ondarenet,[6] l'arxiu electrònic del patrimoni digital basc. Des del 2009, la Biblioteca Nacional d'Espanya (BNE) encarrega captures periòdiques del domini .es a l'Internet Archive, amb seu als Estats Units.

Malgrat la imperfecció de la major part dels sistemes informàtics que serveixen a les polítiques nacionals de preservació del patrimoni digital en la xarxa, l'arxiu d'Internet és actualment una realitat arreu del món tecnològicament desenvolupat.

El PADICAT, el Patrimoni Digital de Catalunya

Amb el canvi de mil·lenni, i igual que la resta de serveis d'informació, les biblioteques nacionals d'arreu d'Europa van iniciar un procés d'evolució del seu model de servei a fi de constituir-se com a equipaments oberts i accessibles, amb

capacitat per donar servei a tots els seus clients potencials, presencials o no. Amb el punt de mira en aquest objectiu, una gran part de les biblioteques nacionals o patrimonials han reorientat la seva direcció estratègica per prendre com a referència els plans estratègics de centres com la British Library o la National Library of Scotland.[7]

Alineada amb aquest corrent europeu i sota la direcció de Dolors Lamarca, la BC va aprovar el 2004 un pla estratègic[8] per evolucionar cap a un model de biblioteca oberta, fiable i orientada a l'usuari, impulsant canvis radicals en la normativa d'accés i préstec, i creant serveis virtuals i projectes digitals amb una doble finalitat: facilitar l'accés obert i universal al coneixement i al patrimoni de Catalunya, i contribuir a preservar-lo.

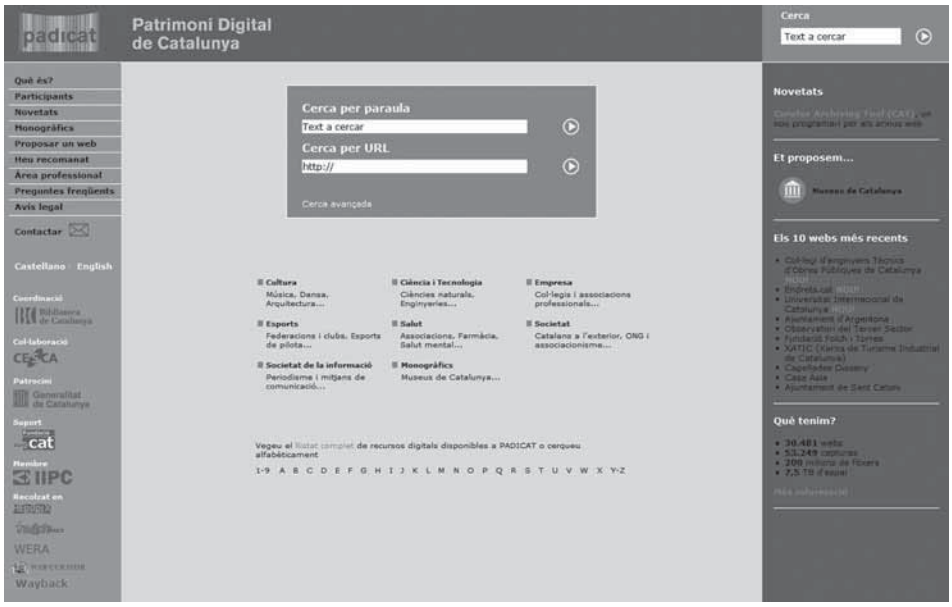
Alguns dels projectes propis o cooperatius fruit d'aquesta determinació[9] són Memòria Digital de Catalunya[10], ARCA (Arxiu de Revistes Catalanes Antigues)[11], RACO (Revistes Catalanes d'Accés Obert)[12], CLACA (Clàssics Catalans)[13], Google Llibres[14] o PADICAT (Patrimoni Digital de Catalunya).

El PADICAT és un dipòsit destinat a compilar, processar i donar accés permanent a la producció digital catalana a Internet. És, sintèticament, l'arxiu web de Catalunya, dedicat a preservar els recursos digitals, essencialment pàgines web, publicats a Internet per al públic de Catalunya.

Coordinat per la Biblioteca de Catalunya, compta amb la col·laboració del CESCA (Centre de Supercomputació de Catalunya) i el finançament de la Generalitat de Catalunya.[15]

A partir d'una fase inicial d'anàlisi dels dipòsits existents,[16] el dipòsit ha experimentat un període de naixement (2005-2006), creixement (2007-2008) i consolidació (2009-2011), que és vigent en el moment de redacció d'aquest article.

Des de l'11 de setembre del 2006 manté operatiu i actualitzat el portal www.padicat.cat, en català, castellà i anglès. Tota la col·lecció és accessible en obert i en línia,[17] consultable per cerca, per navegació a directori temàtic, o per accés directe a paquets monogràfics.



Gràfic 1. **Portada www.padicat.cat.**

D'acord amb el model híbrid, tendència generalitzada en repositoris similars, la política de col·lecció del dipòsit es basa en les accions següents:

- Compilar massivament els recursos digitals publicats en obert a Internet, per mitjà de la captura del domini .cat.[18]
- Impulsar el dipòsit sistemàtic de la producció web de les entitats i les empreses de Catalunya, mitjançant la identificació i la signatura d'un conveni de cooperació.[19]
- Promoure línies de recerca per mitjà de la presentació temàtica dels recursos digitals capturats relatius a determinats esdeveniments de la vida pública catalana, com ara campanyes electorals a Internet, el fenomen de la música en línia, o els museus a Internet.[20]

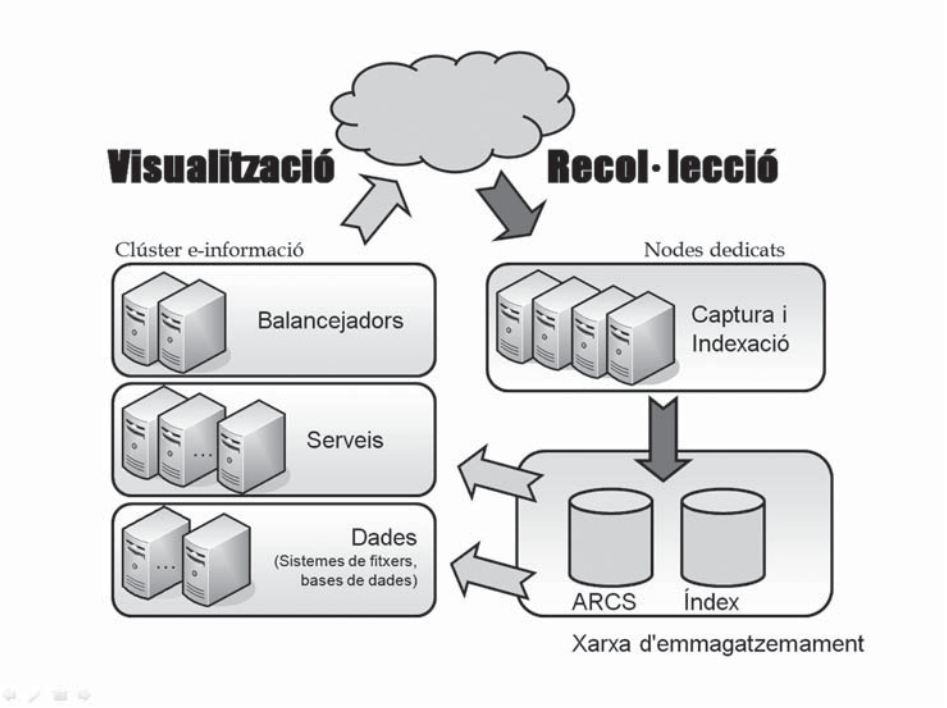
Després de cinc anys d'existència, el dipòsit conté 118.587 captures de 39.587 pàgines web i està format per 249 milions de fitxers informàtics, amb una mida de 7,5 TB.[21]

Aspectes tècnics

Pel que fa a l'arquitectura tècnica del sistema, posteriorment a la fase d'anàlisi i test de programari es va determinar que s'utilitzaria el programa informàtic Heritrix[22], emprat en la major part de projectes de captura de recursos digitals.

Aquest és el programa encarregat de compilar les pàgines web tal com les veu l'usuari que navega per Internet i emmagatzemar-les en arxius comprimits en format ARC[23]. A continuació, el programari Heritrix es complementa amb NutchWax[24], o bé la combinació d'Hadoop[25] i Wayback[26], que duen a terme uns processos d'indexació de la informació compilada que permeten, ulteriorment, utilitzar aquests índexs per localitzar els recursos dins de la col·lecció mitjançant les seves respectives interfícies de consulta: Wera[27], que permet la cerca per paraules clau a través dels índexs generats per NutchWax; i Wayback, que permet la consulta directa per URL en els índexs generats per Hadoop i el mateix Wayback.

Finalment, s'ha aprofitat el programa Web Curator Tool[28], desenvolupat per la National Library of New Zealand i la British Library, com a sistema de gestió documental que permet l'assignació de metadades a una part significativa de la col·lecció, amb la intenció de poder integrar, en el futur, els fons del dipòsit a la cerca en altres catàlegs, tant de la Biblioteca de Catalunya com d'altres institucions.



Gràfic 2. **Arquitectura del PADICAT.**

D'altra banda, el personal del CESCA, soci tecnològic del projecte, ha desenvolupat i compartit amb la comunitat diverses aplicacions *ad hoc*, com els mòduls del CAT (Curator Archiving Tool), dissenyats per millorar l'accés i la recuperació dels recursos digitals dipositats al PADICAT.[29] Tot el programari emprat és de codi obert i gratuït.

Pel que fa al maquinari que sosté el sistema, es compta amb sis nodes HP ProLiant DL360 G4p, encarregats de les tasques de recol·lecció i indexació de les pàgines web. De la cerca i la visualització de resultats en la interfície web, se n'encarrega un clúster Linux d'alta disponibilitat amb característiques de balanceig de càrrega de peticions i de tolerància d'errors en cas de desastre tècnic dels nodes que integren la plataforma. Una cabina NetApp FAS3170 presenta un espai de disc via NFS a aquests nodes. El sistema es completa amb un robot on es conserven còpies de seguretat de les dades en cinta i, en el moment de redacció d'aquest article, un dipòsit de preservació digital a llarg termini desenvolupat per la BC es troba en fase pilot.

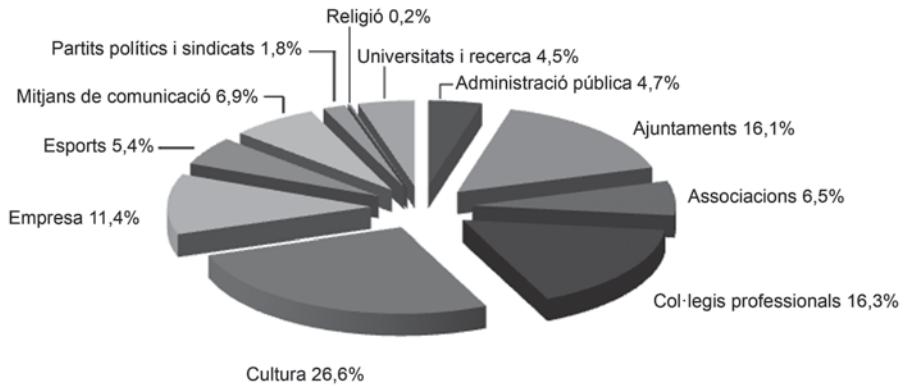
Un cop consolidada la infraestructura tècnica, la previsió de creixement anual per al 2011 s'estableix en 75.700 noves captures d'unes 32.000 pàgines web.

Aspectes legals

Des del plantejament inicial del dipòsit, les limitacions legals han estat analitzades amb rigor però també amb lògica. Malgrat l'obsolescència del text legal espanyol relatiu al dipòsit legal,[30] la llei vigent dóna cobertura a la Biblioteca de Catalunya per a la formació de dipòsits digitals de pàgines web.[31] De fet, països com Suïssa o els Països Baixos han creat i mantenen arxius web sense ni tan sols tenir lleis de dipòsit legal. En tot cas, els països preservadors són lluny de la magnífica legislació danesa, que permet a la seva biblioteca nacional capturar qualsevol web publicada pel públic danès.

Més enllà dels condicionants legals, ja analitzats a peu de nota, la BC ha compartit la defensa d'una filosofia explotada amb èxit des del 1996 per l'Internet Archive, segons la qual la captura de les parts públiques d'Internet és bàsica per preservar la cultura i el patrimoni de la nostra comunitat, igual que han fet les biblioteques amb els llibres, les revistes, els discos i les pel·lícules que al llarg del temps hi ha hagut. Com ha indicat Vives[32], les administracions i els professionals disposem d'arguments bons i suficients per convèncer els nostres dipositants de la bondat dels repositoris, sense entrar en debats estèrils sobre la legalitat o no de preservar la producció digital.

Partint d'aquesta seguretat i en compliment de la política de col·lecció basada en els agents productors de les pàgines web a Catalunya, la BC ha signat 450 convenis de cooperació amb entitats i empreses de tots els sectors, que formalment li permeten capturar, processar i preservar les captures fetes dels seus recursos digitals i donar-hi accés obert.



Gràfic 3. **Tipologia de les entitats participants al PADICAT.**

Preservació digital

La BC és conscient de l'oferta d'estratègies més habituals de preservació,[33] com la migració periòdica o *refresh* de les dades (migració cap a noves versions dels mateixos programes o llenguatges, o cap a nous programes capaços de llegir els anteriors); l'emulació (l'ús de programari, especificacions, etc., que simulin el moment de la creació), i la recreació (simulació per enginyeria inversa o altres mètodes).

L'estat de la qüestió, en l'àmbit mundial i per a aquest tipus de dipòsits, no preveu grans avenços en la garantia absoluta de preservació, malgrat que la previsió sobre el tipus de fitxers que el repositori ha de gestionar, basada en la composició actual de la col·lecció, reveli que la major part dels fitxers corresponen a formats estàndards, que en les macroxifres poden simplificar en el futur la tasca preservadora. Així, sobre una radiografia de la web catalana, basada en una mostra de 226 milions de fitxers, el 95% correspon a estàndards suscepti-

bles de ser accessibles en el futur per mitjà de migracions massives: text/HTML (84%), imatge JPEG o GIF (10%), PDF (1%).[34]

Tipus	Fitxers	
text/HTML	190.153.788	84,27%
image/JPEG	19.019.039	8,43%
image/GIF	4.648.446	2,06%
application/PDF	3.165.429	1,40%
image/PNG	1.453.780	0,64%
application/atom+XML	932.301	0,41%
application/RSS+XML	847.729	0,38%
text/XML	814.114	0,36%
text/plain	706.356	0,31%
application/x-shockwave-Flash	517.686	0,23%
text/CSS	478.121	0,21%
no-type	357.020	0,16%
text/DNS	320.236	0,14%
application/x-JavaScript	268.239	0,12%
application/octet-stream	266.803	0,12%
application/MSMord	203.847	0,09%
image/TIFF	187.962	0,08%
image/PJPEG	153.988	0,07%
audio/MPEG	118.540	0,05%
Altres	1.045.872	0,46%
Total	**225.659.296**	

Gràfic 4. **Tipologia dels fitxers dipositats al PADICAT, basant-se en una mostra de 226 milions.**

Grau de compliment de les expectatives

A començament del 2006 es van fer públics els objectius del PADICAT per al període 2006-2011,[35] i se'n va publicar un primer balanç l'any 2008.[36]

En principi, s'establia que l'objectiu genèric del PADICAT era dissenyar i produir un sistema que permetés que la BC compilés, processés i donés accés permanent a la producció digital catalana. Cinc anys després, l'objectiu genèric del projecte s'ha traduït efectivament en el disseny i la producció d'un sistema

que ens permet actualment compilar, processar i donar accés a la part de la producció digital catalana a Internet que hem incorporat al dipòsit.

En un marc més operatiu, en la planificació del projecte s'assenyalaven tres eixos de treball que continuen vigents, atès que són característics dels models híbrids de captura. A continuació s'especifica el grau de compliment d'aquests objectius:

Compilar massivament els recursos digitals publicats en obert a Internet. A partir d'una sèrie de captures de prova (2007-2008) del domini .cat, que van obligar a ampliar sensiblement els recursos destinats a captura i emmagatzematge, s'ha dut a terme una captura exhaustiva del domini corresponent a l'any 2009, dues més per a l'any 2010,[37] i està programada la captura sistemàtica semestral. Per tant, les pàgines web amb domini .cat es capturen i processen anualment. Complementàriament, es fa una captura semestral dels recursos procedents de les entitats que han signat convenis de col·laboració; una captura semestral dels recursos digitals procedents de recomanacions,[38] i captures periòdiques de recursos que formen part dels monogràfics.[39]

Impulsar el dipòsit sistemàtic de la producció web de les entitats i les empreses de Catalunya. Des de l'inici del projecte, i amb l'objectiu de tancar 500 convenis de cooperació abans del final del 2011, s'han identificat fins a 2.000 institucions considerades agents principals de la producció digital catalana. S'ha presentat el projecte a 1.800 d'aquests ens, i s'han formalitzat els 450 convenis de cooperació, amb una previsió per als propers mesos de complir l'objectiu de 500 entitats.

Promoure línies de recerca per mitjà de la presentació temàtica dels recursos digitals sobre determinats esdeveniments de la vida pública catalana. A partir de l'anàlisi de processos similars en altres projectes, i coincidint amb un calendari electoral regular, s'ha optat per efectuar una captura focalitzada d'un esdeveniment electoral anual relacionat amb campanyes electorals: al Parlament de Catalunya el 2006, les municipals del 2007, al Congrés i al Senat espanyol el 2008, al Parlament Europeu el 2009, i novament al Parlament de Catalunya el 2010. Una acció de col·laboració amb l'Escola Superior de Música de Catalunya (ESMUC) va permetre ampliar aquesta oferta amb una nova fórmula: els recursos digitals catalans relacionats amb la música folk-rock. Complementàriament, s'ha presentat un monogràfic dedicat als museus de Catalunya[40] i s'han fet captures ràpides d'esdeveniments a Internet, com el seguiment del debat de la prohibició dels toros al Parlament, l'editorial «La dignitat de Catalunya»,

«Zona9 música a la xarxa», les prèvies de la campanya Jocs Olímpics Barcelona 2022 o els casals catalans a l'exterior.

En la planificació del projecte s'assenyalaven vuit objectius complementaris als tres principals que s'acaben de descriure. Aquest n'és el balanç del grau de compliment:

Creació d'una xarxa de contactes del projecte que garanteixi el suport institucional i permeti difondre l'acció de la BC en el seu territori de referència. A part del CESCA, soci tecnològic imprescindible, i la Fundació puntCAT, soci privilegiat del programa, 1.800 entitats de tot tipus han estat contactades en nom de la direcció de la Biblioteca de Catalunya i s'ha pogut explicar el projecte segons un circuit de treball predefinit; 450 d'aquestes entitats han signat un conveni de col·laboració amb el projecte, i moltes altres estan en diverses fases del procés que finalitza amb la signatura.

Posició de la BC en una situació de lideratge pel que fa a la preservació digital de pàgines web. A Espanya, el projecte PADICAT va ser pioner. Amb l'arribada dels projectes Ondarenet i les captures del domini .es per part de la BNE s'han establert tímides línies de cooperació. En l'àmbit internacional, el projecte forma part de la principal xarxa de treball en preservació digital, l'International Internet Preservation Consortium (IIPC). La BC, d'altra banda, ha assistit durant aquests sis anys a un centenar d'actes professionals per explicar el projecte, amb la qual cosa ha projectat una imatge de lideratge en preservació del patrimoni digital i ha tingut diversos impactes en mitjans de comunicació especialitzats i generalistes a partir de l'emissió periòdica de comunicats de premsa i altres fórmules comunicatives.

Aprenentatge per part de la BC dels líders mundials en preservació digital. El projecte, a remolc de l'objectiu anterior, es troba en situació d'aprendre de les entitats internacionalment pioneres: l'Internet Archive, les biblioteques nacionals escandinaves, els grups de treball d'aquests organismes, etc. La distància física i la llengua de contacte, en tot cas, no permeten aprofitar sinergies (projectes idèntics amb objectius similars arreu del món) en la mesura que es podria desitjar. Les llistes de distribució i les reunions esporàdiques no supleixen qualitativament, encara, les possibilitats d'aprenentatge mutu. D'altra banda, els projectes internacionals d'arxiu web que estan consolidats no dediquen a aquestes tasques els recursos que caldrien per a la millora permanent de les seves eines, millora de la qual es podrien nodrir projectes com el PADICAT.

Creació d'una eina que permeti capturar, processar i oferir en obert els recursos digitals que formen el patrimoni digital de Catalunya. La provisió

d'equips de maquinari i de personal expert per part del soci tecnològic, el CESCA, ha permès produir un instrument que compleix aquesta necessitat basant-se en la utilització del programari que ja s'emprava en altres projectes. Aquest instrument ha estat descrit en l'arquitectura i els aspectes tècnics del PADICAT. Tanmateix, ha estat i continua sent una tasca complexa disposar d'una eina eficaç a l'hora de garantir aquest procés bàsic, especialment pel que fa a la recuperació necessària dels documents capturats, utilitat en què presenta més errades el programari dels arxius web coneguts.

Provisió d'accés obert i en línia als recursos dipositats. El 2006 es va inaugurar la web del PADICAT en una versió trilingüe que avui es manté. Tal com s'ha descrit, com a filosofia de projecte s'ha donat accés obert via Internet a tota la col·lecció disponible. Primer, amb el motor de cerca a text complet. En una segona fase, amb la creació de centres d'interès monogràfics. Finalment, es van completar les opcions anteriors amb la creació d'un directori temàtic, dedicat als públics que prefereixen la navegació com a fórmula de visita dels fons que formen el dipòsit.

Creació d'un sistema de posicionament per metadades aplicable a la interfície de cerca. Tenint en compte el rol que exerceix la BC en la normalització de les eines que permeten una descripció bibliogràfica correcta i la catalogació de documents de tota mena, el projecte PADICAT va apostar per catalogar, a través d'un sistema estàndard de metadades, el màxim nombre de recursos digitals dipositats. Per mitjà de l'externalització dels processos de catalogació, s'han aplicat metadades estàndard al 30% de la col·lecció. Està en desenvolupament[41] l'eina que permetrà reflectir de manera automàtica les metadades en el sistema de posicionament del motor de cerca del dipòsit. Tenim els recursos correctament catalogats, però encara no podem utilitzar aquesta informació per millorar el sistema de cerca i recuperació del dipòsit, ni tampoc integrar els resultats en altres catàlegs, objectiu final de la catalogació per metadades.

Traç de les línies de la futura preservació digital de pàgines web de Catalunya. Ningú no dubta que la preservació correcta dels recursos digitals és un gran repte de la nostra societat. El projecte que ens ocupa ha desvetllat la radiografia de formats de la web catalana com a pas bàsic per projectar polítiques de preservació. Complementàriament, el PADICAT ha format part del grup de treball de preservació digital de la BC, que ha definit les característiques i les funcionalitats del repositori de preservació que és en fase de desenvolupament a la Biblioteca de Catalunya.[42]

Previsió que al final del 2011 el volum del dipòsit contingui unes 100.000 pàgines web capturades en diverses edicions. El dipòsit conté actualment 118.039 captures, de 39.587 pàgines web, i està previst un creixement exponencial durant el 2011, atès que s'ha assolit la infraestructura tècnica. Les xifres s'especifiquen en el proper capítol, dedicat als reptes de futur.

A mode de conclusió, i el més important, sens dubte, és que s'està fent satisfactòriament un treball sistemàtic de compilació, processament i difusió del patrimoni digital de Catalunya a Internet.

Reptes de futur

El futur del PADICAT, després d'unes etapes que considerem de naixement (2005-2006), creixement (2007-2008) i consolidació (2009-2011), passa per sistematitzar la seva capacitat de creixement, per millorar els seus processos de treball i per optimitzar els recursos de què disposa.

La fita numèrica anual, a partir del gener del 2011, és avançar en els objectius descrits, per mitjà de la incorporació al dipòsit d'unes 75.700 versions d'aproximadament 32.000 pàgines web:

- 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.
- Compilació semestral dels 800 recursos procedents de recomanacions.
- Compilació única dels 1.000 recursos de les eleccions municipals del 2011.
- Compilació diària d'una part substancial de 30 publicacions seriades en línia.

Pel que fa a l'estratègia de futur, en primer lloc cal **consolidar i garantir la infraestructura** necessària del projecte, adequant-la als objectius del sistema, o bé modificar a la baixa aquests objectius. L'estructura actual de maquinari i de personal expert en el programari que s'utilitza permet treballar amb la capacitat necessària per abordar el repte de la captura global de la web catalana, però un decreixement causat per l'obsolescència dels recursos tècnics comportaria, lògicament, la paràlisi de l'arxiu web.

En segon lloc, és imprescindible abordar la definició de les **estratègies de preservació digital** dels fitxers que conté el dipòsit que ens ocupa. Probablement sigui un dels aspectes clau en el retorn que la BC vol fer a la societat. A banda de radiografies periòdiques de la web catalana, que il·lustren la diagnosi

del llenguatge de programació que s'usa en l'edició digital, el sistema pot ajudar a definir quins formats experimenten, a curt termini, problemes d'il·legibilitat. I a partir de constatar aquestes pèrdues, és possible traçar cap a quins formats cal transformar els fitxers per dotar-los de dosis més elevades de permanència, a més dels processos que han de fer possible aquesta transformació.

En tercer lloc, el PADICAT ha de continuar apostant per l'eix de treball que ha tingut més impacte en l'ús que han fet els mitjans de comunicació i també des dels estudis universitaris especialitzats en les respectives matèries: l'impuls de línies de recerca a partir de la **creació de col·leccions monogràfiques**. Com ha esdevingut norma, és profitós reforçar aquestes accions amb la implicació d'experts que assessorin la BC en la identificació dels recursos digitals que podem considerar de referència.[43]

En quart lloc, tal com s'ha apuntat en els objectius numèrics, la **creació de l'hemeroteca digital a Internet** és un repte destacat. L'abordatge de la captura sistematitzada de publicacions en sèrie a Internet s'ha treballat al llarg del 2010 per projectar les necessitats infraestructurals de l'acció, que s'inicia el gener del 2011.

En cinquè lloc, malgrat l'estandardització dels llenguatges informàtics que s'empren en el programari del PADICAT i la resta de projectes similars, cal destacar que no és encara possible, com és d'esperar, un intercanvi eficaç de registres bibliogràfics, a fi de poder integrar tots els dipòsits existents, o aquests dipòsits en altres catàlegs. L'ús de passarel·les i llenguatges estàndards és encara en fase d'implementació en el programari del projecte que, insistim, és comú en la majoria de dipòsits digitals com el PADICAT. De la capacitat d'incidir en el **desenvolupament del programari que permeti l'intercanvi de registres** depèn també la consecució dels objectius de futur de la BC, en la seva voluntat d'arxivar la web catalana.

Finalment, és essencial impulsar la **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.

Notes

1 *Guidelines for the preservation of digital heritage*. Canberra: UNESCO, 2003. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf> [Consulta: 15, desembre, 2010]. Hi ha una versió en castellà.

2 Per a una panoràmica global sobre aquests projectes vegeu: LLUECA, C. «Webs sempre accessibles: les biblioteques nacionals i els dipòsits digitals nacionals». *BiD: textos universitaris de biblioteconomia i documentació*. Núm. 15 (desembre, 2005). <http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec2.htm> [Consulta: 15, desembre, 2010], i també un recorregut esquemàtic pels arxius web associats a l'IIPC a: «Member archives». *International Internet Preservation Consortium*. <http://netpreserve.org/about/archiveList.php> [Consulta: 15, desembre, 2010].

3 Una part important dels projectes, incloent-hi el PADICAT, es troben representats en l'International Internet Preservation Consortium (IIPC): <http://netpreserve.org/> [Consulta: 15, desembre, 2010].

4 El portal PADICAT, http://www.padicat.cat és operatiu des del 2006.

5 Interessant reflexió sobre comunitats nacionals a Internet a: GOMES, D.; SILVA, M. J. «Characterizing a National Community Web». *ACM Transactions on Internet Technology*. Vol. 5, núm. 3 (agost 2005). <http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf> [Consulta: 15, desembre, 2010].

6 El portal Ondarenet, http://www.ondarenet.kultura.ejgv.euskadi.net és operatiu des del 2007.

7 Vegeu en el cas britànic: BRITISH LIBRARY. *Redefining the Library.* London: BL, 2004. <http://www.bl.uk/aboutus/stratpolprog/strategy0811/blstrategy20052008.pdf> [Consulta: 15, desembre, 2010]; i en l'escocès: National Library of Scotland. *Breaking through the walls.* Edinburgh: NLS, 2003. <http://www.nls.uk/media/22415/strategy2004.pdf> [Consulta: 15, desembre, 2010].

8 A partir del que consta en les lleis catalanes de biblioteques del 1981 (DOGC 123, 29/04/1981) i 1993 (DOGC 1727, 29/03/1993), la Biblioteca de Catalunya té com a missió recopilar, conservar i difondre la producció bibliogràfica catalana i la relacionada amb l'àmbit lingüístic català, a més, vetlla per la conservació i la difusió del patrimoni bibliogràfic. Entenem que aquest patrimoni bibliogràfic inclou també la producció bibliogràfica digital publicada a Internet. Vegeu els documents d'estratègia: Biblioteca de Catalunya. *Pla estratègic de la Biblioteca de Catalunya 2004-2008*. Barcelona: BC, 2004. <http://www.bnc.es/bc-/qualitat/pestrategic2004_2008.doc> [Consulta: 15, desembre, 2010]; i Biblioteca de Catalunya. *Pla estratègic de la Biblioteca de Catalunya 2009-2012*. Barcelona: BC, 2009. <http://www.bnc.es/bc/qualitat/pestrategic_2009_2012.pdf> [Consulta: 15, desembre, 2010].

9 L'estratègia i els projectes, a excepció del Google Llibres, que va ser un acord posterior, foren presentats a: LAMARCA, D.; SERRA, E. «L'estratègia de la Biblioteca de Catalunya en projectes digitals». *Ítem*. Núm. 41 (set.-des. 2005). P. 41-43. <http://www.raco.cat-/index.php/Item/article/view/40866/68116> [Consulta: 15, desembre, 2010].

10 Format per quinze institucions, i coordinat per la Biblioteca de Catalunya i el Consorci de Biblioteques Universitàries de Catalunya, Memòria Digital de Catalunya és el dipòsit digital que exerceix de paraigua a altres repositoris, i incorpora tot tipus de documentació. La BC hi ha incorporat 80.000 documents. Disponible a <http://mdc.cbuc.cat/> [Consulta: 15, desembre, 2010].

11 El mèrit és doble: completar les col·leccions repartides per institucions diverses; i impulsar-ne la digitalització per garantir accés i preservació. Actualment inclou 250 títols disponibles a <http://www.bnc.cat/digital/arca/index.html> [Consulta: 15, desembre, 2010].

12 RACO és un repositori cooperatiu del Consorci de Biblioteques Universitàries de Catalunya, del Centre de Supercomputació de Catalunya i de la Biblioteca de Catalunya, en el qual participen 57 institucions, i en què es poden consultar, en accés obert, els articles a text complet de revistes científiques, culturals i erudites catalanes. Disponible a <http://www.raco.cat/> [Consulta: 15, desembre, 2010].

13 CLACA és un servei de centres d'interès entorn de personalitats culturals i erudites catalanes, que fa accessible documentació diversa digitalitzada relacionada amb cada autoritat. Disponible a <http://www.bnc.cat/fons/claca.php> [Consulta: 15, desembre, 2010].

14 Conegut mundialment com Google Books, a Catalunya en formen part la Biblioteca de Catalunya, que n'és la coordinadora, la Biblioteca de l'Ateneu Barcelonès, la Biblioteca Pública Episcopal del Seminari de Barcelona; la Biblioteca del Centre Excursionista de Catalunya, i la Biblioteca de l'Abadia de Montserrat. L'objectiu és possibilitar l'accés lliure a Internet de la còpia digital de 200.000 documents lliures de drets, alhora que es preserven digitalment aquests valuosos documents en les respectives seus. Accessible a <http://books.google.cat/> [Consulta: 15, desembre, 2010].

15 El pressupost del projecte és d'uns 800.000 euros des del 2005. L'aportació per al trienni 2009-2011 és de 584.711,99 euros (impostos inclosos), i comprèn la infraestructura tecnològica i el personal de desenvolupament i manteniment informàtic, a més del personal de coordinació i manteniment de la col·lecció i del portal web.

16 Vegeu la memòria de l'anàlisi i el plantejament del projecte: Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: BC, desembre 2005. <http://www.recercat.net/handle/2072/1757> [Consulta: 15, desembre, 2010].

17 Des de gener del 2007, i en una trajectòria ascendent, el portal ha rebut una mitjana mensual de 1.277 visites.

18 La Biblioteca de Catalunya va signar, el novembre del 2006, un conveni de cooperació amb la Fundació puntCAT per accedir als registres amb domini .cat.

19 En el moment de redacció d'aquest article, 450 institucions, detallades més endavant per sectors. Vegeu-ne la llista actualitzada a: <http://www.padicat.cat/participants.php> [Consulta: 15, desembre, 2010].

20 Més endavant se'n detalla la tipologia. Inclou els monogràfics de les campanyes electorals al Parlament de Catalunya el 2006, les municipals del 2007, al Congrés i el Senat espanyol el 2008, al Parlament Europeu el 2009, i novament al Parlament de Catalunya el 2010. Complementàriament, música folk-rock i museus de Catalunya. Vegeu-ho a: <http://www.padicat.cat/esdeveniments.php> [Consulta: 15, desembre, 2010].

21 Dades amb data 15 de desembre del 2010. Vegeu-ne les actualitzacions a: <http://www.padicat.cat/es/estadistiques.php> [Consulta: 15, desembre, 2010].

22 Heritrix: <http://crawler.archive.org>. Vegeu-ne un article explicatiu a: MOHR, G. (*et al.*). «An introduction to Heritrix: an open source archival quality web crawler». *International Web Archiving Workshop*. (2004). <http://www.iwaw.net/04/Mohr.pdf> [Consulta: 15, desembre, 2010].

23 Arc File Format: <http://en.wikipedia.org/wiki/ARC_(file_format)> [Consulta: 15, desembre, 2010].

24 NutchWax: <http://archive-access.sourceforge.net/projects/nutch/> [Consulta: 15, desembre, 2010].

25 Hadoop: <http://hadoop.apache.org/core/> [Consulta: 15, desembre, 2010].

26 Wayback: <http://archive-access.sourceforge.net/projects/wayback> [Consulta: 15, desembre, 2010].

27 Wera: <http://archive-access.sourceforge.net/projects/wera/> [Consulta: 15, desembre, 2010].

28 Web Curator Tool: <http://webcurator.sourceforge.net/> [Consulta: 15, desembre, 2010].

29 Vegeu-ne un article descriptiu a: LLUECA, C. (*et al.*). «CAT (Curator Archiving Toll): millorant l'accés als arxius web». *International Internet Preservation Consortium meeting*. (Viena: 2010). <http://www.recercat.net/bitstream/2072/85525/2/Padicat_iipc_2010_CAT.pdf> [Consulta: 15, desembre, 2010].

30 Malgrat que l'actualització sigui imminent, el text legal espanyol vigent data del 1971. És un exemple de bona pràctica la Llei danesa del dipòsit legal, del 2004. Vegeu-ne una traducció a l'anglès a: *Act on legal deposit of published material: translation of Act N. 1439 of 22 December 2004: unauthorized version*. <http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html> [Consulta: 15, desembre, 2010], i un exhaustiu informe comparatiu a: GEORGIA, A. *Digital legal deposit in the EU member states: an overview of regulatory and implementation status:*

background report in connection on digital libraries. Frankfurt: Foundation Conference of European National Librarians, 2006. <http://web3.nlib.ee/cenl/docs/Digital%20Legal%-20Deposit%20in%20the%20EU%20member%20states.pdf> [Consulta: 15, desembre, 2010].

31 Essencialment pel que fa a la captura, l'emmagatzematge i la transformació dels fitxers inclosos en les pàgines web capturades.

32 VIVES, J. «Aspectos de propiedad intelectual en la creación y gestión de repositorios institucionales». *El profesional de la información*. Vol. 14, núm. 4 (juliol-agost 2005). <http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/4.pdf> [Consulta: 15, desembre, 2010].

33 AYRE, C.; MUIR, A. «The right to preserve: the rights issues of digital preservation». *D-Lib magazine.* Vol. 10, núm. 3 (març 2004). <http://www.dlib.org/dlib/march04/ayre/03ayre.html> [Consulta: 15, desembre, 2010].

34 Accediu a radiografies actualitzades a: <http://www.padicat.cat/es/estadistiques.php> [Consulta: 15, desembre, 2010].

35 LLUECA, C. «El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya». *10es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 2006. <http://eprints.rclis.org/archive-/00006434/01/llueca_padicat.pdf> [Consulta: 15, desembre, 2010].

36 CÓCERA, D.; LLUECA, C. «PADICAT: realitat i reptes de 3 anys d'arxiu web de Catalunya». *11es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 2008. <http://eprints.rclis.org/archive/00013562-/01/llueca_padicat_jornades_2008.pdf> [Consulta: 15, desembre, 2010].

37 El domini .cat tenia 31.125 registres actius durant la darrera captura massiva, el 2010, que es completen a efectes numèrics amb pàrquings, readreçaments i pàgines sense contingut, d'acord amb les dades públiques de la Fundació puntCAT.

38 El projecte promou la participació activa de l'usuari per mitjà de la recomanació de webs susceptibles de formar part de l'arxiu. Aquesta possibilitat, oberta a través d'un formulari, ha tingut un èxit considerable pel que fa a la participació dels usuaris (758 pàgines recomanades, en la data de presentació d'aquest article). No ha passat el mateix, però, en la rapidesa a l'hora de procedir a la captura d'aquests recursos, ja que s'han produït retards en el procés de captura i publicació.

39 L'any 2010 es van capturar de manera focalitzada 1.000 pàgines web per a la campanya electoral al Parlament a Internet, amb un volum aproximat de 4.000 captures.

40 Selecció i captures de 1.532 recursos digitals relatius als 657 museus i col·leccions museogràfiques de Catalunya. Vegeu el monogràfic a: <http://www.padicat.cat/museus.php> [Consulta: 15, desembre, 2010].

41 El programari CAT (Curator Archiving Tool), esmentat en la nota 29. Els tests realitzats indiquen que el sistema de retorn basat en el programari WERA prioritza, de més a menys ponderació, els elements principals següents: l'URL del recurs; un conjunt de paraules properes al terme o els termes de cerca (context); el nom del domini web (cal remarcar la diferència entre l'URL, que indica la ruta sencera del document web, i el domini, que és el nom principal del lloc web del qual «penja» el document); el títol del document web i, finalment, la frase resultant, acotada per signes de puntuació, dins la qual hi ha el terme o els termes cercats. No obstant això, les expectatives del projecte són incidir en un posicionament òptim dels recursos donant prioritat a les metadades que els catalogadors del PADICAT adjunten als recursos capturats (paraules clau de matèria, títol normalitzat, etc.).

42 Accediu a «Repositori de preservació de la Biblioteca de Catalunya: informe descriptiu i de situació», publicat el desembre del 2010 al dipòsit RECERCAT.

43 Accions d'aquest estil s'han dut a terme amb professors universitaris en la selecció de recursos de les campanyes electorals, i el resultat ha estat molt satisfactori, entenem que per a ambdues parts.

RESUM

El PADICAT (Patrimoni Digital de Catalunya) és el dipòsit digital creat el 2005 per la Biblioteca de Catalunya amb l'objectiu de capturar, processar i donar accés permanent al patrimoni digital de Catalunya a Internet. Fonamenta la seva política de col·lecció en una estratègia híbrida, basada en la captura massiva del domini .cat, la captura selectiva per conveni dels agents productors de les pàgines web catalanes i la captura focalitzada d'esdeveniments públics. El repositori ofereix el seu fons en obert a Internet. Després de cinc anys d'experiència, i a partir del context internacional, aquest article descriu el sistema de funcionament i els reptes de futur del PADICAT.

RESUMEN

El PADICAT (Patrimonio Digital de Cataluña) es el depósito digital creado en el 2005 por la Biblioteca de Cataluña con el objetivo de capturar, procesar y dar acceso permanente al patrimonio digital de Cataluña en Internet. Fundamenta su política de colección en una estrategia híbrida, basada en la captura masiva del dominio .cat, la captura selectiva por convenio de los agentes productores de las páginas web catalanas y la captura focalizada de acontecimientos públicos. El repositorio ofrece su fondo en abierto en Internet. Después de cinco años de experiencia, y a partir del contexto internacional, este artículo describe el sistema de funcionamiento y los retos de futuro del PADICAT.

ABSTRACT

PADICAT (Digital Heritage of Catalonia) is a digital repository created in 2005 by the Library of Catalonia with the aim of collecting, processing and providing permanent access to the digital heritage of Catalonia via the Internet. Its collection policy is founded on a hybrid strategy, which is based on the massive collection of the .cat domain, selective collection by agreement with the agents producing Catalan web pages and collection focused on public events. The archive collection is freely accessible on the Internet. Five years down the line and with the international context in mind, this article describes the PADICAT operating system and the future challenges that await it.

RÉSUMÉ

Le PADICAT (patrimoine numérique de Catalogne) est le dépôt numérique créé en 2005 par la bibliothèque de Catalogne dans le but de capturer, traiter et donner un accès permanent au patrimoine numérique de Catalogne via Internet. Sa politique de collection est fondée sur une stratégie hybride, qui se base sur la capture massive du domaine.cat, la capture sélective sur convention des agents

producteurs des pages Internet catalanes et la capture focalisée d'événements publics. Le référentiel offre son fonds en accès ouvert sur Internet. Après cinq ans d'expérience et à partir du contexte international, cet article décrit le système de fonctionnement et les enjeux futurs du PADICAT.