PADICAT: realitat i reptes de 3 anys de l'arxiu web de Catalunya

Daniel Cócera
Ciro Llueca
Projecte PADICAT (Patrimoni Digital de Catalunya)
Biblioteca de Catalunya
Hospital, 56 - 08001 Barcelona
Tel. 932 702 300 (ext. 2129)
padicat@bnc.cat

Resum

El PADICAT, projecte de la Biblioteca de Catalunya per arxivar la web catalana, treballa en tres eixos d'actuació (captures temàtiques selectives, captures exhaustives, acords amb els agents productors), ha obtingut resultats d'implementació òptims en les aliances de cooperació amb 300 institucions i empreses del país; en diverses accions temàtiques de captures selectives; i en presència en entorns professionals internacionals; així com en diverses accions de captura sistematitzada de recursos digitals publicats a Internet. D'altra banda, en els processos de captura exhaustiva i especialment en la cerca i visualització de la informació processada es troben les mancances més evidents d'un sistema que, arreu del món, no és encara una realitat consolidada. La Biblioteca de Catalunya traça les accions previstes en preservació digital de les pàgines web, així com en normalitzar el que ha de convertir-se en un sistema eficaç de conservació del patrimoni digital, que doni garanties plenes als productors dels recursos digitals de Catalunya.

Paraules clau: Dipòsits digitals; Biblioteques nacionals; Preservació digital; Arxius web.

1. Introducció

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. Tal com ho exposen les *Directrices para la preservación del patrimonio digital*,[1] els recursos que són fruit del coneixement o l'expressió dels éssers humans, ja siguin de caràcter cultural, educatiu, científic o administratiu, o comprenguin informació tècnica, jurídica, mèdica i d'un altre tipus, es generen cada cop més sovint directament en format digital, o es converteixen a aquest format a partir de material analògic ja existent.

Així plantejava el seu inici la comunicació de maig de 2006 que presentava a la comunitat bibliotecària el projecte PADICAT (Patrimoni Digital de Catalunya) en la desena edició de les Jornades Catalanes d'Informació i Documentació,[2] que organitza bia-

1. *Directrices para la preservación del patrimonio digital*. Canberra: Unesco, 2003. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 25/01/2008]

2. Llueca, C. (2006). «El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya», *10es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliote-

nualment el Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya. A partir de constatar la consolidació del format digital en la nostra societat, i des de la dècada dels noranta —continuava la contribució a les jornades— diverses biblioteques nacionals han promogut estratègies per garantir l'accés permanent a la producció digital: la preservació de les pàgines web. L'administració catalana, amb el lideratge de la Biblioteca de Catalunya (BC, http://www.bnc.cat) i el suport tecnològic del Centre de Supercomputació de Catalunya (CESCA, http://www.cesca.cat), s'havia afegit a aquest propòsit amb el projecte PADICAT el juny de 2005. Fa ara tres anys.

El primer any de projecte, el 2005, es va dedicar íntegrament a la planificació del sistema.[3] Des de llavors, i complint les previsions de la Biblioteca de Catalunya, el dipòsit digital ha normalitzat les seves línies d'actuació i tanmateix ha pogut conviure amb la realitat de la problemàtica diària en la captura, processament, i difusió de recursos digitals publicats a Internet. Paral·lelament, el projecte ha entrat a formar part de la comunitat internacional en matèria de preservació digital de llocs web, il·lustrant que el sistema empès a Catalunya es troba en la primera línia mundial en aquesta matèria. Addicionalment, s'assumeix que la provisional manca generalitzada d'una plataforma tecnològica adequada ha causat que moltes de les accions previstes no han pogut desenvolupar-se tan eficaçment com es preveia.

La present contribució vol mostrar a la comunitat professional catalana l'estat dels avenços en la implantació del projecte PADICAT després de tres anys del seu inici. Els autors volen compartir els nivells de compliment dels objectius proposats, en una tasca que és inèdita a Catalunya, i a Espanya, així com els reptes de present i de futur que es plantegen. Finalment, és objectiu de la contribució traçar les perspectives de futur immediat en el projecte que lidera la Biblioteca de Catalunya, per tal de fer públiques les previsions en preservació digital de les pàgines web creades a Catalunya.

2. Arxivant la Web

L'arxiu de la Web, com popularment es coneix el conjunt de tècniques dirigides per la creació de dipòsits digitals com el PADICAT, no és avui una funció consolidada arreu de les biblioteques nacionals o entre altres òrgans competents en preservació patrimonial. Existeixen diversos dipòsits nacionals en funcionament.[4] Els més coneguts, com ja s'ha

caris-Documentalistes de Catalunya, 2006. <http://eprints.rclis.org/archive/00006434/01/llueca_padicat.pdf>. [Consulta: 25/01/2008].

3. Resultat d'aquesta acció és la publicació de: Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: Biblioteca de Catalunya, 2005. <http://www.recercat.net/handle/2072/1757> [Consulta: 25/01/2008].

4. Per a una panoràmica recent vegeu Llueca, C. (2005). «Webs sempre accessibles: les biblioteques nacionals i els dipòsits digitals nacionals». *BiD: textos universitaris de biblioteconomia i documentació*, núm. 15 (des 2005). <http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec1.htm> [Consulta: 25/01/2006].

fet notar en escrits anteriors, són també els que el 1996 van iniciar els respectius Webs nacionals: el suec *Kulturarw3* i l'australià *Pandora*, així com un dipòsit d'abast internacional, el gegant *Internet Archive*.

Sabem positivament, arrel dels diversos contactes mantinguts amb institucions paral·leles a la BC, que podem comptar fins a 36 projectes en diverses fases de funcionament, essent accions consolidades, amb funcionament regular, un terç d'aquesta xifra.[5] Igualment, podem afirmar que el model híbrid és el més generalitzat: consisteix a compilar recursos digitals en base a la captura exhaustiva de recursos web, tot combinant aquestes accions automatitzades amb processos de selecció més biblioteconòmics, focalitzats en un tema determinat, o a partir de la selecció manual de recursos.

Lamentablement, el nombre de dipòsits que, com el PADICAT, permeten accedir obertament a les seves col·leccions o fons, és molt limitat.[6] Sovint es tracta d'evitar potencials conflictes amb la vulneració dels drets de propietat intel·lectual dels recursos capturats sense autorització expressa, però complementàriament a aquesta legítima precaució, un factor determinant és el fet que no s'hagin perfeccionat les interfícies de recuperació de la informació dipositada.

En tot cas, hem d'assenyalar que les plataformes tecnològiques emprades per aquests tipus de dipòsits estan molt centrades en els tres aspectes bàsics de la cadena documental: la captura de recursos; la indexació d'aquests recursos un cop capturats; i l'accés als recursos emmagatzemats. Però en menor mesura en la preservació d'aquests recursos, terreny que és pràcticament verge si l'associem exclusivament als processos dirigits a garantir l'accés permanent a les pàgines web capturades, i no, com passa en la majoria dels casos, si s'analitza des d'un punt de vista més global (preservació de materials digitalitzats, preservació de fitxers ofimàtics, etc.).

3. El projecte PADICAT, 2005-2008

3.1. Grau de compliment dels objectius principals

Com s'especificava a bastament en la comunicació de presentació en societat del PADICAT,[7] a partir de la missió de la Biblioteca de Catalunya establíem l'objectiu genèric

5. Un bon indicador és el llistat dels 36 membres de l'International Internet Preservation Consortium (IIPC, http://netpreserve.org), si suposem que les entitats que estan disposades a pagar una quota anual tenen, almenys, un projecte en marxa, i la intenció de compartir el seu coneixement amb la resta. A Espanya, tindrem notícia en els propers mesos de projectes d'arxiu web que inicien el seu camí, complementant així la tasca que executa en solitari el PADICAT.

6. L'australià Pandora (http://pandora.nla.gov.au/), el britànic UK Web Archive (http://www.webarchive.org.uk/), el japonès WARP (http://warp.ndl.go.jp/), l'Internet Archive (http://www.archive.org), i el PADICAT (http://www.padicat.cat).

7. Llueca, C. (2006). «El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya», *10es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliote-

del programa que ens ocupa en dissenyar i produir un sistema que permeti a la BC compilar, processar i donar accés permanent a la producció digital catalana.

A tres anys del seu inici, 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 que hem incorporat al dipòsit. De la permanència d'aquesta informació se'n parlarà més endavant. Però destaquem que la creació d'aquest sistema, avui consolidat, garanteix a la BC, i per tant al país, comptar amb una eina eficaç que construeix, aquí i ara, l'arxiu web de Catalunya.

En la planificació del projecte, a un nivell més operatiu, s'assenyalaven els tres eixos de treball, que segueixen vigents doncs són característics dels models híbrids de captura. S'especifica el grau de compliment d'aquests objectius:

— Compilació massiva dels recursos digitals publicats en obert a Internet. El projecte PADICAT va signar el novembre de 2006 un conveni de cooperació amb la Fundació puntCAT per tal d'accedir als 18.000 registres sota domini. CAT.[8] Però més enllà de diverses captures parcials d'aquests registres, no s'ha procedit a maig de 2008 a un procés de captura massiva d'aquestes pàgines web, essent la causa principal la capacitat limitada dels recursos destinats a captura i emmagatzemament de les pàgines web. Sí es preveu, en els propers mesos, un reforçament d'aquestes accions exhaustives. Grau de compliment: 10% (diverses captures de 500 recursos, incloent-hi els recomanats pel públic de la web[9]).
— Impuls del dipòsit sistemàtic dels agents implicats en la producció digital a Catalunya. Des de l'inici del projecte la BC, i amb l'objectiu de tancar 300 convenis de cooperació abans del final de 2008, 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 300 convenis de cooperació,[10] essent la previsió per als propers mesos d'augmentar suficientment aquesta xifra. Grau de compliment: 100% (diverses captures de 300 recursos).
— Promoció de línies de recerca específiques per mitjà de la integració focalitzada de recursos digitals sobre determinats esdeveniments de la vida pública catalana. A partir de l'anàlisi de processos similars en altres projectes, i coincidint amb

caris-Documentalistes de Catalunya, 2006. <http://eprints.rclis.org/archive/00006434/01/llueca_padicat.pdf>. [Consulta: 25/01/2008].

8. Accediu a la nota de premsa per a més informació: «Signat el conveni de cooperació entre la Biblioteca de Catalunya i la fundació puntCAT per la preservació de les pàgines web». <http://www.padicat.cat/novetats2006.php#13-11-06>[Consulta 25/01/2008].

9. 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 (400 demandes l'any 2007), no així, però, en la rapidesa a l'hora de procedir a la captura d'aquests recursos, havent-se produït retards en el procés de captura i publicació.

10. Llistat dels participants: http://www.padicat.cat/participants.php [Consulta: 25/01/2008].

un calendari electoral intens en els darrers anys, s'ha optat per realitzar una captura focalitzada de tres esdeveniments (un per any) relacionats amb campanyes electorals: al Parlament de Catalunya 2006, municipals 2007, i al Congrés i Senat espanyol 2008.[11] Tanmateix, un acció de col·laboració amb l'Escola Superior de Música de Catalunya (ESMUC, http://www.esmuc.net/), ha permès ampliar aquesta oferta amb una nova fórmula: els recursos digitals catalans relacionats amb la música folk-rock. En els propers mesos, nous centres d'interès completaran aquest eix de treball. Grau de compliment: 100% (diverses captures de 300 recursos relacionats amb aquests esdeveniments).

Per concloure, és inevitable assenyalar en aquesta serena anàlisi de realitat del projecte que és precisament en els aspectes més automatitzables (captura exhaustiva del .CAT) on no s'assoleix el compromís adquirit amb la BC per part de la direcció del projecte, essent la causa principal la lògica capacitat limitada de la infraestructura tecnològica, i addicionalment la voluntat d'experimentar durant la fase que ens ocupa amb processos de més risc, com són la captura dels recursos procedents de les entitats susceptibles de formar el cos d'agents productors del patrimoni digital, la captura de recursos relacionats amb les campanyes electorals: un nombre limitat de recursos capturats amb una alta periodicitat (diària, o setmanal) en detriment de captures massives semestrals d'un nombre elevat de recursos.

Per als propers mesos podem esperar un grau de compliment global que estimem del 70%, a banda de la satisfacció d'haver efectivament creat un sistema de funcionament ja consolidat, que ha estat reconegut a nivell internacional.[12]

2.1. Grau de compliment dels objectius complementaris

Complementàriament als tres objectius genèrics esmentats en l'apartat anterior, establim el grau de compliment dels objectius secundaris:

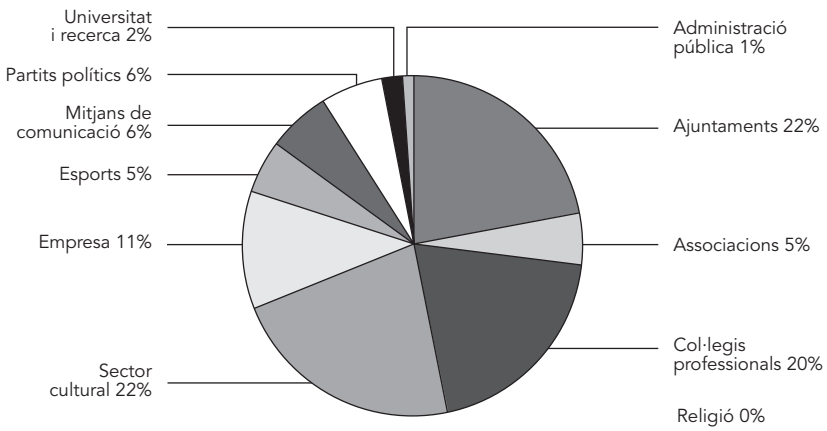
— Creació d'una xarxa de contactes del projecte que garanteixi el suport institucional i permeti difondre l'acció de la BC en el seu territori de referència. A part

11. El 23 de gener de 2007, PADICAT presenta i incorpora al seu web el recull especial dedicat a les eleccions autonòmiques de 2006 (http://www.padicat.cat/eleccions2006.php), les dades globals del qual es poden consultar a: http://www.padicat.cat/novetats2007.php#23-01-07. Així mateix, el 23 de novembre de 2007 és presentat el recull especial sobre les eleccions municipals del mateix any (http://www.padicat.cat/municipals2007.php), un producte molt més ambiciós, degut a la pròpia naturalesa atomitzada de l'esdeveniment electoral, les xifres i estadístiques del qual es poden consultar al resum de les dades: http://wwwpadicat.cat/docs/Especial_eleccions_municipals2007.pdf.

12. Més informació a la nota de premsa: «La BC se suma al Consorci Internacional de Preservació d'Internet». <http://www.padicat.cat/novetats2007.php#19-02-07> [Consulta: 25/01/2008].

de 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. 300 d'aquestes entitats (63 entitats culturals; 61 ajuntaments; 59 col·legis i associacions professionals; 31 empreses; 19 mitjans de comunicació; 16 entitats esportives; 8 partits polítics i sindicats; etc.) han signat un conveni de col·laboració amb el projecte, i moltes altres estan en diverses fases del procés que finalitza amb la signatura. Grau de compliment: 70%

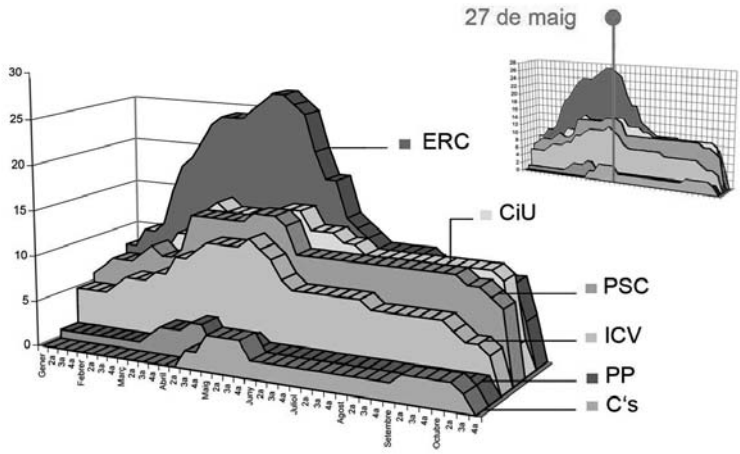
Gràfic 1. Tipologia de socis del projecte



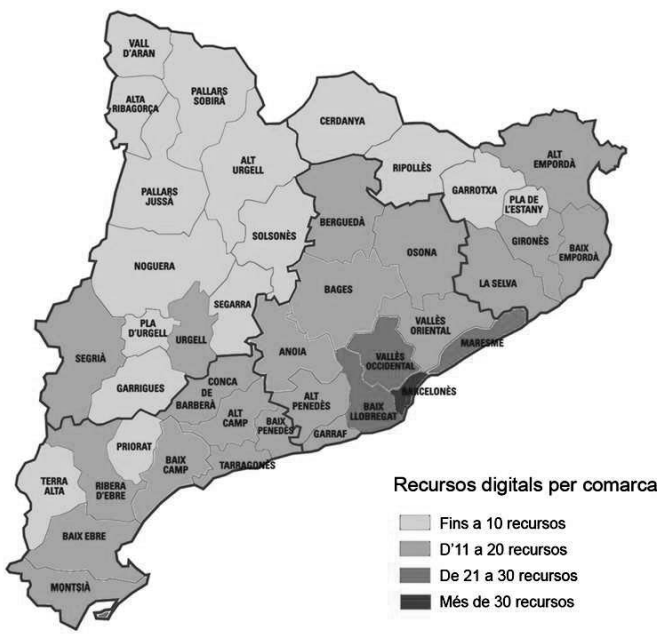
— Posició de la BC en una situació de lideratge pel que fa a preservació digital de pàgines web. A Espanya el projecte PADICAT és pioner, i encara únic. A nivell internacional, el projecte forma part de la principal xarxa de treball en preservació digital, l'Internacional Internet Preservation Consortium (IIPC, http://netpreserve.org), i ha estat distingit per la Library of Congress, responsable de comunicació d'aquest consorci, com a exemple d'arxiu web per les seves accions en les campanyes electorals.[13] La BC, d'altra banda, ha assistit en aquests tres anys a un centenar d'actes professionals per explicar el projecte, projectant una imatge de lideratge en preservació del patrimoni digital, i ha tingut diversos impactes en mitjans de comunicació especialitzats, i també generalistes, a partir de l'emissió periòdica de comunicats de premsa i altres fórmules comunicatives. Grau de compliment: 80%

13. Accediu a la campanya electoral de les eleccions municipals 2007 a: http://www.padicat.cat/municipals2007.php, així com a l'informe resum de les dades: http://www.padicat.cat/docs/Especial_eleccions_municipals2007.pdf [Consulta d'ambdós recursos: 25/01/2008].

Gràfic 2. Índex d'abandonament dels blocs dels candidats en les eleccions municipals 2007



Gràfic 3. Distribució comarcal dels recursos digitals en la campanya de les eleccions municipals 2007



— Aprenentatge de la BC per part 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 l'aprofitament de sinergies (projectes idèntics amb objectius similars arreu del món) en la mesura del que es podria desitjar. Les llistes de distribució i les reunions esporàdiques no supleixen qualitativament, encara, les possibilitats d'aprenentatge mutu. La inexistència de projectes similars a Espanya no ha possibilitat compartir experiències en un entorn que treballa cooperativament en d'altres matèries comunes. D'altra banda, som conscients que els projectes internacionals de dipòsit digital nacional que estan consolidats no dediquen a aquestes tasques els recursos que caldrien per la millora permanent de les seves eines, millora de la que es podrien nodrir projectes com el PADICAT. Grau de compliment: 45%

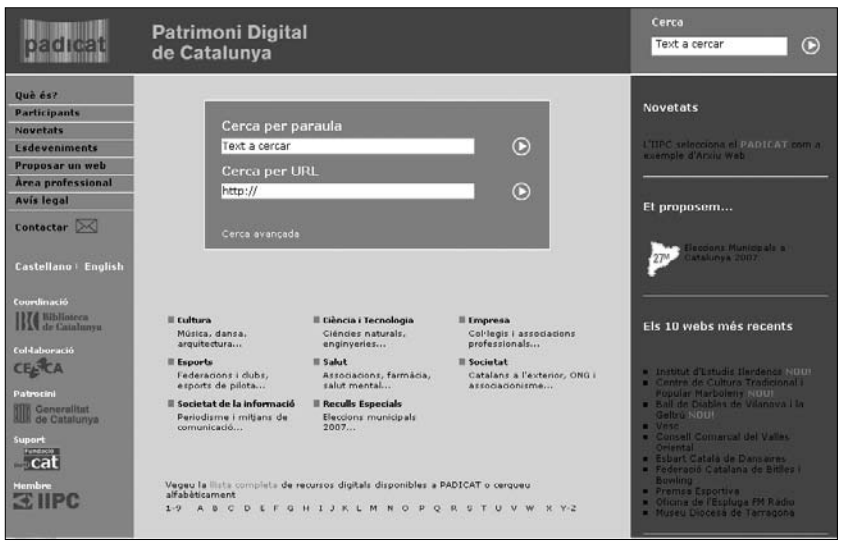
— Creació d'una eina que permeti capturar, processar i oferir en obert els recursos digitals que formen el patrimoni digital de Catalunya. La provisió d'equips de maquinari i de personal expert per part del soci tecnològic, el CESCA, ha permès produir un instrument que compleix aquesta necessitat en base a la utilització del programari que ja s'utilitzava en altres projectes. Nogensmenys, ha estat i segueix sent una tasca complexa comptar amb una eina del tot eficaç a l'hora de garantir aquest procés bàsic, especialment pel que fa a la necessària recuperació impecable dels documents capturats. Per treballar còmodament, el robot PADICAT necessitaria sovint comptar amb uns recursos que superen les necessitats actuals d'altres tipus de dipòsits més consolidats (TDX, Recercat, RACO, etc.). La limitació dels recursos disponibles, com no podria ser de cap altra manera pel pressupost que gestiona el projecte,[14] ha reduït en moltes ocasions l'ambició dels responsables del PADICAT. En positiu, com ja hem remarcat, existeix actualment una eina en estat operatiu que permet la captura de pàgines web, el processament de les mateixes, i l'accés en obert als recursos dipositats en el PADICAT. De la futura provisió de recursos dependrà la seva expansió. Grau de compliment: 60%

— Provisió d'accés obert i en línia als recursos dipositats. L'11 de setembre es va inaugurar la web del PADICAT (http://www.padicat.cat) en una versió trilingüe, que avui es manté. Des del primer dia, 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: paquets temàtics (campanyes electorals a les diverses eleccions, bàsicament) construïts a

14. El pressupost del projecte 2006-2008 ascendeix a 766.000 euros, accediu a la comunicació resultat de l'Acord de govern de gener de 2006: http://www.gencat.net/acordsdegovern/20060124/02.htm [Consulta 25/01/2008].

partir del fons del dipòsit, que permeten arribar a públics específics (per exemple, a professors i estudiants universitaris de ciència política i sociologia, o de comunicació política[15]). Finalment, s'ha completat 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 PADICAT. Més enllà de l'èxit d'aquesta operació, hem de reflectir que els processos de posicionament dels resultats en resposta a cerca a text complet, i molt especialment la presentació dels recursos digitals capturats, són encara lluny del que hom qualificaria d'òptims, a causa bàsicament de la lentitud en l'aparició de les pàgines web dipositades. Grau de compliment: 70%

Gràfic 4. Portada de la web del projecte PADICAT



— 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 la correcta descripció bibliogràfica, així com en la catalogació

15. Per a la identificació i selecció de recursos en la confecció dels monogràfics dedicats a les eleccions, s'ha comptat amb la col·laboració i l'assessorament dels experts en ciència política: professora Rosa Borge (Universitat Oberta de Catalunya), professora Mariona Ferrer (Universitat Pompeu Fabra), professor Miquel Peralta (Universitat Ramon Llull) i professora Mariona Tomàs (Universitat de Barcelona). Aquests professors, i d'altres del mateix àmbit, varen ser contactats després de la publicació de l'especial eleccions autonòmiques 2006 per a realitzar una avaluació sintètica dels recursos. Els seus consells i propostes de millora han servit de guia per a la confecció dels reculls monogràfics posteriors, i alhora, aquests, han esdevingut una important eina de treball per als estudiants d'aquests camps.

de documents de tota índole, el projecte PADICAT va apostar amb fermesa per catalogar, per mitjà d'un sistema estàndard de metadades, el major nombre de recursos digitals dipositats. Actualment quatre catalogadors proveïts per l'empresa Indra formen part de l'equip PADICAT, i aquesta xifra és bona mostra d'aquesta intenció. El cert és que la tasca d'aquest personal no es pot veure encara reflectida en els processos de recuperació de la informació, quan s'accedeix per mitjà del cercador disponible a la web del projecte. Malgrat que s'hi destinen limitadament recursos i personal especialitzat del CESCA, no s'ha identificat la via adequada per incidir en el page rank del programari,[16] per tal d'afavorir una correcta recuperació, més enllà de la cerca a text lliure o per navegació en el directori de recursos. A peu pla: tenim els recursos correctament catalogats però no podem encara utilitzar aquesta informació per millorar el sistema de cerca i recuperació del dipòsit. Grau de compliment: 30%.

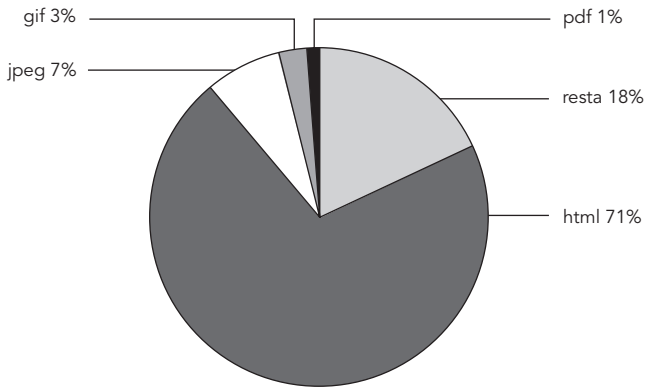
— Traç de les línies de la futura preservació digital de pàgines web de Catalunya. Que la correcta preservació dels recursos digitals és un gran repte per al patrimoni digital de les nostres societats no ho dubta ningú[17] des del 2003, quan la Unesco ho va posar de manifest. La realitat és que el projecte que ens ocupa ha dedicat insuficients energies a explorar les possibilitats en matèria de transformació de fitxers capturats per garantir-ne la correcta permanència, més enllà del contacte habitual amb els experts catalans, i el fet molt destacable de desvetllar quina és la radiografia dels formats de la Web catalana:[18] el 72% dels fitxers ho són de format Html; el 7% de Jpeg; el 2,4 de Gif; l'1,3% de Pdf. La resta representa el 17,5%. Aquestes dades ens diuen dues coses: que la majoria de programadors utilitzen formats coneguts; i que hem de projectar que els formats coneguts seran globalment transformables a futurs codis estàndard, que generalment seran llegibles sense massa dificultat. Grau de compliment: 30%

16. Els tests que han realitzat els analistes del CESCA indiquen que, actualment, el sistema de retorn basat en el software WERA, prioritza, de major a menor ponderació, els següents elements principals: la URL del recurs; conjunt de paraules properes al terme o termes de cerca (context); el nom del domini web (cal remarcar la diferència entre la 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, per últim, la frase resultant, acotada per signes de puntuació, dins la qual hi ha el terme o termes cercats. No obstant, les expectatives del projecte són les d'incidir en un posicionament òptim dels recursos a través de prioritzar les metadades que els catalogadors de PADICAT adjunten als recursos capturats (paraules clau de matèria, títol normalitzat, etc.).

17. Ho constata el fet que als estudis universitaris s'hagi inclòs la formació en preservació de recursos digitals. Tanmateix, cal saludar l'aparició de la primera obra tècnica monogràfica en castellà i català: Keefer, A.; Gallart, N. (2007). La preservació de recursos digitals: el repte per a les biblioteques del segle XXI. Barcelona: UOC.

18. Més informació a la nota de premsa: «L'Html predomina a la web catalana»: http://www.padicat.cat/novetats2007.php#20-12-07 [Consulta: 25/01/2008].

Gràfic 5. Radiografia dels formats més habituals d'una mostra de la Web catalana



Taula 1. Formats més habituals d'una mostra de la Web catalana

Tipus i format del fitxer	Nombre de fitxers	Volum en GB	% del total de fitxers	% del volum total
text/html	24.429.679	592	72	56
image/jpeg	2.416.055	124	7	12
image/gif	834.019	7	2,4	0,6
application/pdf	449.983	167	1,3	16
no-type	75.070	0,2	0,2	0,1
image/png	72.905	1,5	0,2	0,1
application/x-shockwave-flash	68.379	6	0,2	0,5
application/msword	42.150	5	0,1	0,50
text/plain	39.962	16	0,1	1,5
text/css	35.668	0,2	0,1	0,1
text/xml	35.583	0,5	0,1	0,1
application/x-javascript	23.882	0,2	0,1	0,1
image/pjpeg	14.514	0,4	0,1	0,1
audio/mpeg	10.319	41	0,1	4
application/atom+xml	10.264	0,1	0,1	0,1
image/bmp	10.202	2	0,1	0,2
audio/x-ms-wma	8.869	26	0,1	2,4
application/download	8.122	0,3	0,1	0,1
application/zip	5.730	11	0,1	1,1
application/xml	5.396	0,1	0,1	0,1
application/vnd. ms-excel	5.222	0,5	0,1	0,1

— Previsió que al final de la present fase del projecte el volum del dipòsit contingui unes 100.000 versions de pàgines web, equivalents a uns 30 TB de volum. En el proper apartat es relatarà l'estat actual del dipòsit. Sí hem d'avançar que el nombre actual de recursos capturats (llocs web) és molt lluny de les ambicioses previsions inicials a causa, com s'ha explicat, de la voluntat de donar prioritat a aspectes de més complexitat (captures molt seguides de pocs recursos) per sobre d'accions més exhaustives, que sens dubte haurien de reforçar la idea inicial d'un creixement exponencial. Addicionalment, el fet és que el creixement exponencial desitjat va invariablement lligat a una acció prèvia de mesura i dimensionament de la infraestructura necessària per fer possible el procés d'una captura global de recursos. Grau de compliment: 10%

3.3. Estat actual del dipòsit

La gestió pressupostària del projecte transcorre sense novetat.[19] Pel que fa als recursos humans, en el projecte hi treballen 7 persones a temps complet, més els equips del CESCA i la BC, que aporten la seva experiència en temes específics del desenvolupament del PADICAT.

Sense comptar les estacions de treball dels membres de l'equip, el projecte manté en funcionament exclusiu 4 servidors ProLiant DL360 G4p, amb un Robot Scalar i2000 i 4 TB de capacitat. Aquesta estructura es recolza en una altra més potent, la del CESCA, que permet dotar de reforços per a determinades tasques de funcionament del PADICAT.

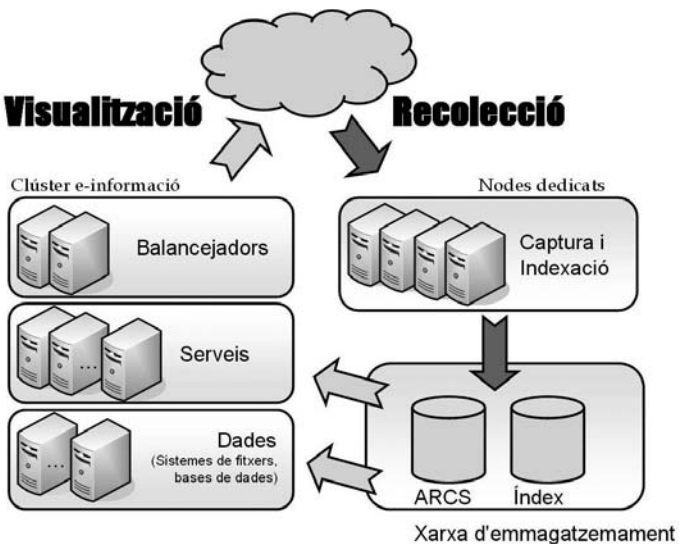
En el moment de la redacció de la present comunicació el dipòsit compta amb unes 3.000 captures d'uns 1.000 llocs web. Cada lloc web ha rebut un tractament diferent (mínim de dues captures). El nombre total de fitxers (html, jpeg, gif…) és de 34 milions, i l'espai que ocupen és de 1,3 TB, malgrat que aquest volum pugui ascendir fins als 4 TB durant determinats processos de treball (indexació, sobretot).

La plataforma digital del projecte, la web del PADICAT, mostra el fons complet de pàgines web (les 3.000 captures d'uns 1.000 llocs web) per mitjà de les eines de cerca ja esmentades: un robot de cerca a text complet, un directori que permet la navegació temàtica, i diversos centres d'interès, a banda d'informes de funcionament, articles i notes professionals, recursos destacats, etc. El lloc web del projecte rep una mitjana mensual de 950 visites.

19. Com s'ha esmentat a la nota 15, el pressupost ascendeix a 766.000 euros. A la memòria del projecte (nota 3), es desglossen les previsions econòmiques.

Gràfic 6. Maquinari del PADICAT a les instal·lacions del CESCA

Gràfic 7. Arquitectura de funcionament del PADICAT



3.4. Reptes immediats

La infraestructura limitada del projecte PADICAT ha fet que al llarg del període analitzat s'hagi treballat amb anticipació d'entre tres i nou mesos, mitjançant calendaris de captura i indexació, els dos processos que alenteixen sensiblement el sistema.

En el moment de redacció de la comunicació que ens ocupa sabem positivament que els mesos de febrer, març, i abril de 2008, estaran dedicats íntegrament a la captura i processament dels recursos digitals de la campanya electoral de les eleccions generals al Congrés i al Senat. Els mesos de maig a juliol estaran dedicats a captura de publicacions seriades, una acció de futur que inicia ara la fase d'implementació. Finalment, els darrers mesos de l'any estaran dedicats a capturar llocs webs de les entitats que tenen convenis de cooperació signats amb la BC.

La interfície de cerca i navegació de la web del PADICAT haurà sofert diversos canvis: implementació del directori temàtic; publicació de nous centres d'interès (incloent-hi el resultant de l'esmentada campanya electoral, etc.); traducció al castellà i a l'anglès d'algunes parts encara monolingües; fusió dels cercadors Wayback i Wera per accelerar la visualització dels fitxers components dels recursos dipositats, etc.

Paral·lelament, és previst dur a terme una anàlisi del sistema de page rank del cercador que utilitza el sistema, per definir amb claredat els paràmetres de posicionament en l'aparició dels resultats a una cerca per text lliure.

4. Reptes de futur

El futur del PADICAT, després d'una etapa que podem considerar de naixement, passa per consolidar la seva capacitat de creixement, per millorar els seus processos de treball, i per optimitzar els recursos de què disposa.

En primer lloc, cal dimensionar la infraestructura necessària del projecte, adequant-la als objectius del sistema, o bé modificar a la baixa aquests objectius. L'actual estructura de maquinari i de personal expert en el programari que s'utilitza no permet treballar amb la capacitat necessària per abordar el repte de la captura global de la web catalana. El fet de comptar amb el CESCA com a soci tecnològic permetrà, de ben segur, establir quines són les necessitats, i en base a aquestes donar resposta tecnològica per al creixement exponencial que perseguim.

En segon lloc, és imprescindible abordar la definició de les estratègies de preservació 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 hom usa en l'edició digital, el sistema pot ajudar a definir quins formats sofreixen, a curt termini, problemes de 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 eleva-

des de permanència, així com dels processos que han de fer possible aquesta transformació.

En tercer lloc, el PADICAT ha de seguir apostant per l'eix de treball que ha resultat de més impacte. La creació de línies de recerca (eleccions, música, etc.) ha donat bons resultats en diversos escenaris, especialment en l'ús que han fet els mitjans de comunicació i també des dels estudis universitaris especialitzats en aquestes matèries. Probablement sigui profitós reforçar aquestes accions destinades a crear centres d'interès amb la implicació de col·lectius d'experts que assessorin la BC en la identificació dels recursos digitals que podem considerar de referència. Accions esporàdiques d'aquest estil s'han realitzat 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.

En quart lloc, l'abordatge de la captura sistematitzada de publicacions en sèrie a Internet és un repte de futur, que s'iniciarà en els propers mesos amb captures que permetran projectar les necessitats infraestructurals del projecte. La revisió del programari existent, per fer possible l'aprofitament dels fitxers ja capturats en les successives captures quan aquestes es repeteixen molt sovint, serà la solució a aquest repte, perquè serà també la manera d'optimitzar els recursos existents.

Finalment, 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 a la majoria de dipòsits digitals nacionals. De la capacitat d'incidir en el desenvolupament del programari depèn també la consecució dels objectius de futur de la BC, en la seva voluntat d'arxivar la web catalana.

En la ja referida comunicació de presentació del projecte es plantejaven els potencials beneficis d'un projecte que es trobava en un estadi preliminar. A tres anys d'aquest inici, els beneficis són plenament vigents, tothora que han esdevingut factors crítics d'èxit en l'estratègia de la BC.

A ulls de la comunitat bibliotecària de Catalunya, els beneficis se centren en la integració dels documents nascuts digitals en la bibliografia nacional, i en el contundent posicionament de la Biblioteca de Catalunya i els socis de projecte en una situació privilegiada com a font d'informació dels documents que representen, en bona mesura, el futur.

Per a la cooperació amb les institucions de la memòria, biblioteques, arxius i museus de Catalunya, així com universitats i centres de recerca, impuls i lideratge en la confecció del patrimoni digital d'Espanya. Tanmateix, la ja existent relació privilegiada amb la resta de biblioteques nacionals del món en termes de preservació digital i dipòsits nacionals.

Per a les institucions, empreses, administracions i particulars que produeixen pàgines web a Catalunya, preservació de la pròpia producció i garantia d'accés, amb els condicionats que la llei regeix, als continguts i dissenys que, d'altra banda, desapareixeran.

Per a la ciutadania, i com es pretén a les directrius de la Unesco, accés obert i permanent als recursos que formen el Patrimoni Digital de Catalunya.

5. Bibliografia

Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: Biblioteca de Catalunya, 2005. <http://www.recercat.net/handle/2072/1757> [Consulta: 25/01/2008]

Biblioteques digitals i dipòsits nacionals de recursos digitals. Barcelona: Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, 1999.

Directrices para la preservación del patrimonio digital. Canberra: Unesco, 2003. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 25/01/2008]

Gomes, D.; Silva, M. J. (2005). «Characterizing a National Community Web». *ACM Transactions on Internet Technology*, vol 5, num 3 (Aug 2005). <http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf>. [Consulta: 25/01/2008]

Hodge, G. M. (2000). «Best practices for digital archiving: an information life cycle approach». *D-LIB Magazine*. Vol. 6, num. 1 (jan 2000). <http://www.dlib.org/dlib/january00/01hodge.html>. [Consulta: 25/01/2008]

Keefer, A.; Gallart, N. (2007). *La preservació de recursos digitals: el repte per a les biblioteques del segle xxi*. Barcelona: UOC.

Llueca, C. (2005). «Webs sempre accessibles: les biblioteques nacionals i els dipòsits digitals nacionals». *BiD: textos universitaris de biblioteconomia i documentació*, núm. 15 (des 2005). <http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec1.htm> [Consulta: 25/01/2008]

Llueca, C. (2006). «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. <http://eprints.rclis.org/archive/00006434/01/llueca_padicat.pdf>. [Consulta: 25/01/2008]

Llueca, C. (2006). «Archivando la Web, el proyecto Padicat (Patrimonio Digital de Cataluña)». *El profesional de la información*. Vol. 15, núm. 6, p. 473-478. <http://eprints.rclis.org/archive/00007767/01/epi_padicat.pdf> [Consulta: 25/01/2008]

Llueca, C. (2007). «Archivando la web catalana, el proyecto PADICAT», *Clip: Boletín de la SEDIC*. Núm. 47. <http://www.sedic.es/p_boletinclip47_confirma.htm>. [Consulta: 25/01/2008]

Torres, N; i altres (2007). «Patrimoni Digital de Catalunya, experiencias del primer año», *Jornadas Técnicas RedIris*. Oviedo: RedIris. <http://www.padicat.cat/docs/poster_padicat_rediris.pdf> [Consulta: 25/01/2008]