El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya

CIRO LLUECA
Coordinador del projecte PADICAT (Patrimoni Digital de Catalunya)
Biblioteca de Catalunya
cllueca@bnc.es

RESUM

Les tecnologies de la informació i la comunicació (TIC) han facilitat que el patrimoni cultural i científic, i la resta d'informació, es presentin en format digital. En resposta a aquest repte les administracions de diversos països, des de la dècada dels noranta, han promogut estratègies per garantir l'accés permanent a la producció digital. Aquesta garantia d'accés passa per assegurar en la mesura de les possibilitats actuals i futures el compliment del repte que suposa el cicle documental clàssic: la compilació, el tractament, la preservació i la difusió de la producció bibliogràfica publicada a Internet: són els *dipòsits digitals nacionals*, nom que reben aquests projectes impulsats habitualment per les biblioteques nacionals. La comunicació descriu el projecte PADICAT (Patrimoni Digital de Catalunya), que la Biblioteca de Catalunya ha posat en marxa per assegurar l'accés permanent a la producció digital catalana.

PARAULES CLAU: Dipòsits digitals, Biblioteques nacionals, Preservació digital, Arxius web

RESUMEN

Las tecnologías de la información y la comunicación (TIC) han facilitado que el patrimonio cultural y científico, y el resto de información, se presenten en formato digital. En respuesta a este reto las administraciones de diversos países, desde la década de los noventa, han promovido estrategias para garantizar el acceso permanente a la producción digital. Esta garantía de acceso pasa por asegurar en la medida de las posibilidades actuales y futuras el cumplimiento de reto que supone el ciclo documental clásico: la compilación, el tratamiento, la preservación y la difusión de la producción bibliográfica publicada en Internet: son los *depósitos digitales nacionales*, nombre que reciben estos proyectos impulsados habitualmente por las bibliotecas nacionales. La comunicación describe el proyecto PADICAT (Patrimonio Digital de Cataluña), que la Biblioteca de Catalunya ha iniciado para asegurar el acceso permanente a la producción digital catalana.

ABSTRACT

Information and communication technologies (ICT) have allowed cultural and scientific patrimony to be presented in digital form —similar to what has occurred with other types of information. In response to the new challenges, since the 1990s the governments of various countries have promoted strategies to guarantee permanent access to their digital content. This guarantee for access includes the assurance, to the extent currently possible, that the traditional document cycle —compiling, handling, preservation and dis-

semination— is maintained for bibliographic material published on the Internet. National digital repositories, the name given to these projects, are commonly driven by national libraries. This paper describes PADICAT (Digital Patrimony of Catalonia) set up by the National Library of Catalonia in order to ensure permanent access to the Catalan production in digital format.

1. Introducció

Les tecnologies de la informació i la comunicació (TIC) 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. Els documents «nascuts digitals», que són creats directament en digital, no existeixen en un altre format que no sigui l'electrònic original, i això en molts casos en detriment de l'ús exclusiu dels formats analògics tradicionals.

En resposta a aquest nou paradigma les administracions de diversos països, des de la dècada dels noranta, han promogut estratègies per garantir l'accés permanent a la producció digital. Aquesta garantia d'accés passa per assegurar en la mesura de les possibilitats actuals i futures el compliment del repte que suposa el cicle documental clàssic, aplicat a la realització de pàgines web: la compilació, el tractament, la preservació i la difusió de la producció bibliogràfica.

Les debilitats i amenaces al repte són notables i han estat ja relatades.[2] Someram ent: l'obsolescència del text legal que possibilita el dipòsit legal a les biblioteques nacionals; el creixement exponencial de la producció digital, sumada a la baixa permanència dels materials publicats a Internet;[3] i el respecte a la legislació en matèria de propietat intel·lectual.

Malgrat les dificultats, diversos països han emprés accions de preservació per assegurar la pervivència de la producció digital. Les biblioteques nacionals han estat sovint les impulsores, i actores principals, d'aquestes accions.

1. *Directrices para la preservación del patrimonio digital*. Canberra: Unesco, 2003. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 25/01/2006]

2. 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]

3. L'UK Web Archiving Consortium fixa en 44 dies la mitjana de vida d'una pàgina web <http://info.webarchive.org.uk/pressrelease21-06-04.html>. [Consulta: 20/01/2006].

El juny de 2005, la Biblioteca de Catalunya va posar en marxa el projecte PADICAT (Patrimoni Digital de Catalunya). L'objectiu d'aquesta comunicació és presentar a les *10es Jornades Catalanes d'Informació i Documentació* la descripció del projecte.

2. Context: els dipòsits digitals nacionals

Un dipòsit[4] digital nacional és l'eina fruit de la iniciativa dedicada a compilar, processar i donar accés als recursos digitals de tot tipus creats en un territori determinat, o sobre aquest territori. Aquests *arxius web* són complementaris als dipòsits institucionals (que es refereixen a la producció digital d'una determinada institució o grup d'institucions), en plena expansió en les societats tecnològicament desenvolupades,[5] o als dipòsits temàtics (amb documents amb una temàtica comuna), més consolidats arreu del món.[6]

Existeixen diversos dipòsits nacionals en funcionament.[7] Els més coneguts són també els que van iniciar-se el 1996 en arxivar la *web nacional:* el suec *Kulturarw3* (http://www.kb.se/kw3/ENG/) i l'australià *Pandora* (http://pandora.nla.gov.au/index.-html); així com un dipòsit d'abast internacional, el gegant *Internet Archive* (http://-www.archive.org). Deu anys més tard, en tot cas, són diverses les iniciatives sorgides amb la mateixa missió i podem comptar fins a 25 projectes en diverses fases de funcionament, essent accions consolidades un terç d'aquest xifra.

4. Dipòsit és la paraula catalana normalitzada que designa el repository anglès.

5. A Espanya, són pioners el *Dipòsit de la Recerca de Catalunya* (RECERCAT, http://www.recercat.net), impulsat pel Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes, la Biblioteca de Catalunya, i el Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya), el *DSpace.Revistes UPC* (https://eprints.-upc.es:8443/revistes) i *DSpace.Eprints UPC* (https://eprints.upc.es:8443/dspace), dedicat a la producció científica de la Universitat Politècnica de Catalunya; o el presentat a l'opinió pública en primera instància: l'*E-Prints* (http://www.ucm.es/eprints) de la Universidad Complutense de Madrid. La resta d'universitats públiques catalanes es troben en disposició de presentar productes similars, en funcionament, en els propers mesos.

6. A Espanya, la *Biblioteca Virtual Miguel de Cervantes* (http://www.cervantesvirtual.com), promoguda per la Fundación del mateix nom, creada per la Universitat d'Alacant, el Grup bancari Santander i la Fundació Marcelino Botín, entre d'altres membres, amb l'objectiu de desenvolupar l'expansió universal de les cultures hispàniques mitjançant la utilització i aplicació de la tecnologia a obres de la literatura, les ciències i la cultura iberoamericana; el portal *Tecnociencia e-revistas* (http://www.tecnociencia.es/-e-revistas), impulsat per la Fundación Española de Ciencia y Tecnología, amb l'objectiu de crear una plataforma digital on es compilin, seleccionin i hostatgin revistes electròniques espanyoles o llatinoamericanes existents amb criteris de selecció qualitatius; o el portal, que no dipòsit, *Temaria* (http://temaria.net), que coordina la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona, dedicat a facilitar la consulta als articles de revistes espanyoles d'Informació i Documentació.

7. 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].

L'anàlisi d'aquestes experiències mostra dos models bàsics de sistemes, amb una tendència generalitzada cap a un model híbrid. El fet que la legislació del dipòsit legal del país en qüestió s'hagi o no actualitzat per permetre legalment la captura de pàgines web per dipòsit legal, no és un tret diferenciador.[8]

El primer és el model integral o exhaustiu (majoritari, i característic dels països escandinaus, entre d'altres), que persegueix la integració automàtica de la web a partir de determinats criteris infrastructurals (lingüístics, segons el domini de les web, segons la ubicació del servidor, etc.).

El segon model és el selectiu (assimilat per Austràlia o el Japó, entre altres països), dedicat a compilar la web en base a una política selectiva (sobre un espai geogràfic determinat, un tema d'interès nacional, etc.).

Aquests dos models han deixat pas, en el que és ja una tendència generalitzada, a models híbrids que complementen la captura periòdica de la web nacional amb accions selectives, ampliant l'abast en alguns casos a determinats esdeveniments d'interès social (eleccions, bàsicament).

El projecte PADICAT és un dipòsit nacional pioner a Espanya.

La Biblioteca de Catalunya té la missió de recollir, conservar i difondre la producció bibliogràfica catalana i la relacionada amb l'àmbit lingüístic català, i vetllar per la conservació i la difusió del patrimoni bibliogràfic.[9] Entenem que aquest patrimoni bibliogràfic inclou també la producció bibliogràfica digital catalana, objecte del sistema que es presenta.

Per tant, prenent com a referent el repte del que parlàvem a la introducció, o la missió que per llei té assignada la Biblioteca de Catalunya, l'estratègia és pertinent: garantir l'accés permanent a la *web catalana.*

3. El projecte PADICAT

3.1. Objectius del projecte

A partir de la missió descrita de la Biblioteca de Catalunya, establim l'objectiu genèric del projecte que ens ocupa, dissenyar i produir un sistema que permeti a la Biblio-

8. El cas danès és paradigmàtic: l'obligació del dipòsit legal cau en el propietari del domini sota el qual el material està publicat, si aquest està assignat específicament a Dinamarca. En relació al material publicat a altres dominis d'Internet, etc., l'editor del material està subjecte a l'obligació del dipòsit legal. És disponible una traducció no oficial de la llei danesa a: http://www.bs.dk/content.aspx?itemguid=%7b332484E6-A5B1-4CEE-B953-059843182050. Per a informació completa sobre el dipòsit legal danès, abans del darrer canvi: Dupont, H. (1999). «Legal deposit in Denmark: the new law and electronic products». *LIBER Quarterly, the journal of European research libraries*, vol. 9, num. 2, (1999). <http://liber-maps.kb.nl/articles/dupont11.htm>. [Consulta: 25/01/2006].

9. Article 7.1 de la «Llei de biblioteques de Catalunya, de 24 d'abril de 1981». *Diari Oficial de la Generalitat de Catalunya*, núm. 123 (29 abril 1981).

teca de Catalunya compilar, processar i donar accés permanent a la producció digital catalana.

L'antecedent del projecte és la ronda de contactes i anàlisi dels sistemes existents que es va realitzar des de la Biblioteca de Catalunya el 1999 per planificar aquest servei.[10]

D'acord amb la tendència generalitzada arreu de les biblioteques nacionals, ja mencionada en l'apartat anterior, el model de dipòsit que persegueix la biblioteca és el sistema híbrid, que consisteix a:

— Compilar massivament els recursos digitals publicats en obert a Internet
— Impulsar el dipòsit sistemàtic dels agents implicats en la producció digital a Catalunya.
— Promoure línies de recerca específiques per mitjà de la integració focalitzada de recursos digitals sobre determinats esdeveniments de la vida pública catalana.

El projecte compta amb la col·laboració del Centre de Supercomputació de Catalunya (CESCA) i el suport del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (DURSI).[11]

3.2. Abast del PADICAT

Cal en primer lloc una definició el més clara possible de quina és la tipologia de recursos tecnològics publicats a Internet, i quines les temàtiques que són susceptibles de formar la col·lecció objecte del sistema.

En abstracte, entenem «Patrimoni Digital» la informació electrònica publicada a Internet, en obert o no, independentment del format en què es presenta aquesta infor-

10. L'aleshores gerent de la BC, Anna Ma. Planet, va documentar aquests esforços a: Planet, A. (1999). «La gestió del dipòsit legal dels recursos digitals catalans i els projectes de col·leccions digitals de la Biblioteca de Catalunya», *Biblioteques digitals i dipòsits nacionals de recursos digitals* (Barcelona: 1999). Barcelona: Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona.

11. En projectes d'aquesta magnitud la col·laboració i la suma d'esforços entre entitats és imprescindible. A Catalunya, els antecedents són els exemples coneguts de cooperació entre institucions: els catàlegs compartits (BEG, CCUC, SLP, XPB, etc. en procés d'unificació); les *Tesis Doctorals en Xarxa* (TDX, http://www.tdx.cbuc.es) i la *Biblioteca Digital de Catalunya* (BDG, http://www.cbuc.es/index5digital.html), impulsades pel Consorci de Biblioteques Universitàries de Catalunya (CBUC, http://www.cbuc.es); el portal de *Revistes Catalanes amb Accés Obert* (RACO, http://sumaris.cbuc.es/raco/quees.html), del mateix consorci, el Centre de Supercomputació de Catalunya (CESCA, http://www.cesca.es) i la Biblioteca de Catalunya (BC, http://www.bnc.es), amb el suport del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (DURSI, http://www.gencat.net/dursi). Dels mateixos impulsors són els dipòsits *Clàssics Catalans* (CLACA) i *Arxiu de Revistes Catalanes Antigues* (ARCA). D'aquest darrer projecte es publica en aquestes mateixes *Actes* una comunicació específica.

mació. Entendrem «de Catalunya» en el sentit que tradicionalment ha tingut la bibliografia nacional de Catalunya en què es basa la política de la nostra biblioteca: tot allò produït a Catalunya o que tracti sobre Catalunya.

3.2.1. Abast tecnològic

La tecnologia que s'aplica als sistemes de dipòsit digital canvia i canviarà en el futur de manera ràpida i sistemàtica, i és evident que les variables sobre la naturalesa del recurs digital, dinamisme, i programari emprat, dota de diferents graus de complexitat al que hom coneix com a *pàgina web,* o directament, *web.*

No entrarem a valorar en profunditat la terminologia usada en relació a les unitats d'informació que representa cada seu web, però sí citarem la definició emprada habitualment pels membres del Laboratori d'Internet del CINDOC-CSIC, que servirà per definir el que genèricament entenem com a web:[12]

> Pàgina web, o conjunt de pàgines web lligades jeràrquicament a una pàgina principal, identificable per una URL i que forma una unitat documental recognoscible i independent d'altres bé per la seva temàtica, bé per la seva autoria, bé per la seva representativitat institucional.

Per tant, entendrem que una web susceptible de formar part de la col·lecció haurà de complir dues condicions bàsiques:

— *Serà una pàgina web identificable per una URL* o un conjunt de pàgines web lligades jeràrquicament a una pàgina principal identificable per una URL
— *Formarà una unitat documental recognoscible*, i independent en grau suficient de la resta per la seva temàtica, autoria, o representativitat institucional.

Possiblement pugui la fase de producció del sistema perfilar concrecions que completin aquesta regla genèrica, així com el tractament que caldrà seguir el procés dels recursos amb unitat pròpia dins d'altres pàgines web,[13] que en un principi no seran tractades independentment de la resta de recursos.

La complexitat pel que fa al tractament de les dades en totes les fases del procés (compilació, emmagatzematge i difusió) s'ha analitzat en profunditat als tre-

12. Interessant reflexió terminològica a: Pareja, V. M. [et al.] (2005). «Desarrollo y aplicación del concepto de sede web como unidad documental de análisis en Cibermetría», *Jornadas Españolas de Documentación (9as: 2005: Madrid)*. Madrid: Fesabid.

13. Alguns exemples aplicables aquí: un gràfic sobre el turisme a Catalunya d'un estudi genèric realitzat a França; una llei d'aplicació a Catalunya en un recull de jurisprudència europeu; la referència al pintor Salvador Dalí en una pàgina web sobre pintors surrealistes, etc.

balls[14] de l'International Internet Preservation Consortium (IIPC, http://www.netpreserve.org), tot establint una classificació que parteix dels documents HTML estàtics (HTML, GIF, JPEG, etc.) i arriba a les aplicacions JavaScript (menús de navegació, informació dinàmica, aplicacions de veu, URLs generades per mecanismes dinàmics, etc.), entre d'altres aspectes.

En conseqüència, i malgrat que la intenció del projecte PADICAT és exhaustiva, la pròpia dinàmica dels sistemes automàtics de captura presenten limitacions en determinades fases d'aquests eixos, com són els canvis molt freqüents, la dependència a la interacció, i especialment l'accés a recursos electrònics d'accés restringit per mitjà de contrasenyes, control d'IPs, etc.

3.2.2. Abast temàtic

Com han recollit alguns autors, Internet està dissenyada per trencar les barreres geogràfiques i fer la informació accessible universalment.[15] Malgrat aquest tret definitori, és possible identificar parts d'aquesta xarxa que continguin mòduls d'interès de grups concrets, als que podem anomenar «comunitats d'usuaris web» i aquestes parts d'interès comú poden ser definides com el grup de documents que es refereixen a certa temàtica o són d'interès de la comunitat.

En l'anàlisi dels projectes existents arreu s'ha reflexionat sobre l'abast temàtic dels dipòsits digitals nacionals. Alguns exemples són:

— El cas australià,[16] on una part significativa del recurs hauria de ser:

- Sobre Austràlia, o
- Sobre un tema de significança i rellevància social, política, cultural, religiosa, científica o econòmica, alhora que està produïda per un autor australià, o
- Escrit per una autoritat australiana reconeguda, alhora que constituir una contribució al coneixement internacional.

— El cas suec,[17] que inclou:

14. Marill, J. [et al.] (2004). *Web harvesting survey*. Version 1 (jul 2004). International Internet Preservation Consortium. <http://netpreserve.org/publications/iipc-r-001.pdf> [Consulta: 25/01/2006] i Boyko, A. [et al.] (2004). *Test bed taxonomy for crawler*. Version 1 (jul 2004). International Internet Preservation Consortium. <http://netpreserve.org/publications/iipc-r-002.pdf>. [Consulta: 25/01/2006].

15. 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: 25/01/2006].

16. «What is the scope of the Archive? Do you have selection guidelines?». *Pandora.* Canberra: Nacional Library of Australia. <http://pandora.nla.gov.au/panfaqs.html#scope>. [Consulta: 25/01/2006].

17. Persson, K. (2005). «Defining Swedish web pages and finding them?». *Kulturarw3*. Stockholm: The Royal Library. <http://www.kb.se/kw3/ENG/Description.htm>. [Consulta: 25/01/2006].

- Les seus web amb domini.se (Suècia) i.nu («ara», en suec i altres llengües escandinaves, molt utilitzat), o
- Les seus web amb dominis internacionals (.com,.org,.net) ubicades a servidors al territori suec, o
- La *Suecana extreana*: les webs que parlen sobre Suècia, viatges per Suècia, o traduccions d'obres literàries sueques.

— El cas danès,[18] que recull en l'article 8.2 de la seva llei de dipòsit legal que el material publicat en xarxes electròniques de comunicació es considera danès quan:

- Està publicat a dominis d'Internet, etc., els quals són específicament assignats a Dinamarca, o
- Està publicat a altres dominis d'Internet, etc., i està adreçat al públic a Dinamarca

— Finalment, el cas portuguès[19] l'estableix en el grup de documents que contenen informació relativa a Portugal o d'interès majoritari de la gent portuguesa, tot considerant la web portuguesa els documents que satisfan les condicions:

- Hostatjat a una seu web sota el domini PT, o
- Hostatjat a una seu web sota el domini.COM,.NET,.ORG, o.TV, escrits en llengua portuguesa i amb almenys un enllaç entrant originat a una web hostatjada en una pàgina amb domini.PT

A partir d'aquests exemples, la informació electrònica susceptible de formar part del PADICAT és el grup de documents creats a Catalunya, o que contenen informació relativa a Catalunya o d'interès majoritari de la gent catalana,[20] aspecte que conceptualment queda ja recollit a l'article 7 de la Llei de biblioteques de 1981:[21]

> La Biblioteca de Catalunya, com a biblioteca nacional, és el primer centre bibliogràfic de Catalunya i té la missió específica de recollir i de conservat tota la producció impresa, sonora i visual, *que s'hi ha produït i s'hi produeix*, per a la qual cosa és la col·lectora del dipòsit legal.

18. *Act in Legal Deposit of Published Material: traslation of Act No. 1439 of 22 December 2004: unauthorized version*. <http://www.bs.dk/content.aspx?itemguid=%7b332484E6-A5B1-4CEE-B953-059843182050>. [Consulta: 25/01/2006]

19. 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: 25/01/2006]

20. De gran complexitat és determinar objectivament l'interès majoritari d'una societat, per les possibles incoherències al llarg del temps.

21. «Llei de biblioteques de Catalunya, de 24 d'abril de 1981». *Diari Oficial de la Generalitat de Catalunya*, núm. 123 (29 abr 1981).

També acull i conserva la producció impresa, sonora i visual, *en català o que fa referència als Països Catalans produïda fora de Catalunya*.

Concretament, establim l'abast temàtic del Patrimoni Digital de Catalunya en la següent estratègia:

— Webs sota domini.CAT,[22] o
— Webs ubicades a servidors de Catalunya,[23] o
— Webs sota dominis geogràfics (.ES,.COM,.NET,.ORG, etc.[24]) en llengua catalana,[25] o
— Webs que no compleixin els requisits anteriors, però relacionades temàticament amb Catalunya[26]

Definit l'abast conceptual del projecte, el sistema de captura, organització i accés ha de contemplar les variables relacionades amb cadascuna de les qüestions que es plantejaran.

22. L'associació puntCAT (http://www.puntcat.org) va presentar la candidatura i impulsar l'aprovació d'aquest domini, dirigit a *la comunitat lingüística i cultural catalana a Internet.* Sense entrar a fer previsions sobre l'ús que es faci del domini, la Biblioteca de Catalunya té en aquest una oportunitat única de recopilar automatitzadament un perfil de recursos relacionats plenament amb la presència catalana a Internet, des de l'inici del seu funcionament. La posada generalitzada en funcionament del domini es va iniciar el gener de 2006, amb la web destinada a promocionar el nou domini: http://www.domini.cat/

23. Per mitjà del control de les seves IP. L'organisme encarregat de l'assignació d'IP és l'Internet Assigned Numbers Authority (IANA, http://www.iana.org), i la seva secció europea és Réseaux IP Européens (RIPE, http://www.ripe.net). RIPE ofereix els serveis d'identificació dels administradors amb direccions IP assignades (una de les variants del que es coneix com servei de WHOIS —*qui és*, en llengua anglesa—). Descartem per inviable, en no oferir el desglossament de les zones europees, el llistat genèric de rang d'IP assignats a Europa (RIPE: 193, 194, 195…) que ofereix el mateix IANA. Però en tot cas molt probablement és a partir d'aquests serveis d'on sorgeixen els paquets que diverses empreses, com *Ip2location* (http://www.ip2location.biz), *Phpcontrol* (http://www.phpcontrol.com) o *Maxmind* (http://www.maxmind.com) ofereixen amb informació sobre IP, a efectes de segments i estudis de mercat. Una altra via per rastrejar IP i serveis és analitzant què passa a la xarxa amb eines com *Netcraft* (http://www.netcraft.com). Finalment, l'agència *Network Information Center para Espanya* (ESNIC, http://www.esnic.es) és una altra fórmula per obtenir llistats dels dominis els organismes dels quals tinguin seu a Catalunya sota domini.ES, que a data de la present redacció són un total de 45.374. Devem un agraïment a les aportacions d'Ana Nistal, Subdirectora de Continguts de la Dirección del Observatorio de las Telecomunicaciones y la Sociedad de la Información de Red.es.

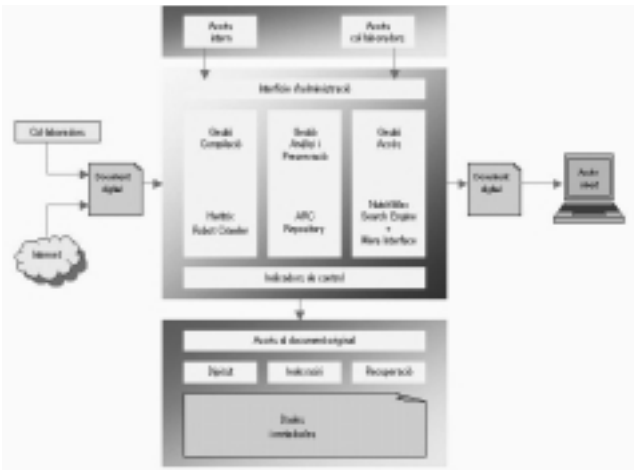
24. A partir del treball de Baeza-Yates, R.; Castillo, C.; López, V. (2005). «Characteristics of the Web of Spain». *Cybermetrics*, Vol. 9, Issue 1, Paper 3 (2005). <http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html>. [Consulta: 25/01/2006]i concretament a http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html#tblInternalDomains, podem apostar per l'odre inicial susceptible de contemplar els dominis de les webs catalanes segons les dades de la Web espanyola:.COM (65%),.ES (16%),.ORG (7,5%),.NET (7%),.INFO (0,8%),.BIZ (0,3%),.TV (0,1%), etc.

25. Sobre la identificació de recursos en llengua catalana devem un agraïment a Joan Soler i Bou, responsable del Diccionari de Freqüències de l'Institut d'Estudis Catalans, així com al professors del Departament d'Estadística i Investigació Operativa de la Universitat Politècnica de Catalunya.

26. A partir de directoris, com Dmoz, Google, Yahoo, etc.

3.3. Funcionament del sistema

Es proposa en el gràfic un model conceptual amb la descripció somera dels aspectes que s'hi mencionen.[27] Només difereix en qüestions puntuals del cicle documental clàssic de les biblioteques i serveis d'informació.



A manca d'una anàlisi futura del sistema informàtic que permet aquest cicle, identifiquem les parts claus del procés en la captura dels recursos, l'organització dels recursos, i l'accés permanent als recursos.[28]

3.3.1. Captura dels recursos

La compilació dels recursos s'orienta en les línies principals[29] que ja s'han apuntat en els objectius de projecte, o sigui:

— Compilar massivament els recursos digitals publicats en obert a Internet

- *Automatitzada* o integralment[30]
- *Manualment*

27. Basat en Dulabahn, B. (2004). «The National Digital Information Infrastructure and Preservation Program (NDIIP): future directions and relevance to other countries». *Archiving web resources*. Canberra: National Library of Australia. <http://www.nla.gov.au/webarchiving/>. [Consulta: 25/01/2006]

28. A tractar específicament en una futura ponència. L'enginyer informàtic Leandro Stasi, de l'empresa Auseba, forma part de l'equip PADICAT des de novembre de 2005.

29. D'acord també amb Christensen-Dalsgaard, B. [et al.] (2003). *Final report for the pilot project «netarkivet.dk»*. Kobenhavn: The Royal Library. <http://netarkivet.dk/rap/webark-final-rapport-2003.pdf>. [Consulta: 25/01/2006]

30. *Snapshot* en llengua anglesa, que podríem traduir com *foto fixa*.

— Impulsar el *dipòsit voluntari* sistemàtic de la producció web dels agents implicats a Catalunya
— Promoure línies de recerca per mitjà de la integració dels recursos digitals de determinats *esdeveniments* de la vida pública catalana.

Cadascuna d'aquestes línies de treball persegueix completar l'estratègia relativa a l'abast temàtic, segons es desprèn del següent quadre, tenint en compte el programari[31] i maquinari[32] que s'han emprat en la fase test del projecte així com les perspectives d'inversió en recursos humans per al projecte pilot.

	Domini.CAT	Servidor a Catalunya	Llengua catalana	Relació temàtica
Automatitzada	X	X	X	
Manual				X
Dipòsit voluntari				X
Esdeveniments				X

A banda, és previsible que el test del programari aporti una base dels conflictes que caldrà superar en el pla pilot (2006), com els problemes relacionats amb el temps de captura, el rebuig per les instruccions del fitxer Robot.txt,[33] l'anàlisi dels *logs* per l'accés denegat, etc. En la referida la fase pilot del projecte s'abordaran aquests conflictes i les vies de resolució.

3.3.2. Organització dels recursos

L'organització dels recursos web ha de permetre gestionar la col·lecció i assegurar-ne la recuperació, alhora que preservar els continguts digitals amb les mesures que la Biblioteca de Catalunya tingui al seu abast.

31. A partir de 1996 Internet Archive va desenvolupar un paquet de programes sota el paraigua del que es coneix com Heritrix (http://crawler.archive.org), al qual s'han anat sumant la resta de projectes, especialment de les biblioteques nacionals escandinaves (Nordic National Libraries) i la Bibliothèque Nationale de France. Els mòduls del programari inclouen Heritrix (capturador), BAT (gestor d'arxius), NutchWax (indexador) i Wera (interfície de consulta). Un article fonamental al respecte és: Mohr, G. [et al.] (2004). «An introduction to Heritrix: an open source archival quality web crawler». *International Web Archiving Workshop* (4th: 2004: Bath). <http://www.iwaw.net/04/proceedings.php?f=Mohr>. [Consulta: 25/01/2006]

32. Dos PCs amb processador Intel Pentium IV 3.2GH, 2GB de RAM i Disc Dur d'1,2 TB (3x400GB)

33. La captura no pot ser obligada segons la legislació vigent, per això el dipòsit ha de ser voluntari i amb acords amb les institucions. Una manera d'evitar la visita dels robots és per mitjà de l'arxiu robot.txt, que fa possible que el robot només arribi a determinades parts (o a cap) de la web. El programari permet la Biblioteca de Catalunya decidir si es respecta o no aquest tipus de limitació. Els precedents són variables: Internet Archive ho respecta escrupolosament. Netarkivet (Dinamarca) no ho respecta mai.

Es proposaran a continuació els detalls relatius a la identificació permanent dels recursos, l'aplicació de metadades, l'emmagatzematge i la preservació.

És previst que en tots els moments del procés l'equip de persones que són administradores del sistema hi tinguin accés per a fer correccions i modificacions. Per contra, cal definir en quines passes del procés hi podria tenir accés el públic extern, especialment en els casos de dipòsit voluntari.

3.3.2.1. Identificació permanent

Com s'ha explicat a bastament (Muxach i Lopo, 1999)[34] la descripció de recursos en la xarxa en basa en l'URI (Uniform Resource Identifier = Identificador Uniforme de Recurs), entenent-lo com la forma que usen els sistemes per a identificar i accedir als fitxers localitzats en els ordinadors connectats. De fet, però, l'URI no és concret, i sí un concepte que n'engloba tres altres que sí representen més que les sigles:

— URL (Uniform Resource Location = Localitzador Uniforme de Recursos), sistema per a localitzar i accedir a un recurs que no distingeix, arquitecturalment, el fitxer de la màquina on és.
— URN (Uniform Resource Name = Nom Uniforme de Recurs), que pretén donar un nom únic i permanent d'un determinat recurs.
— URC (Uniform Resource Characteristic = Característica Uniforme de Recurs), que permet incloure metadades sobre un determinat recurs.

La tendència inicial, apuntada per aquests autors el 1999, era que la majoria de col·leccions continuessin emprant l'URL en referir-se a la localització del document digital. Després d'una revisió dels treballs del CBUC[35] i les previsions terminològiques que entenem vigents, es decideix per la denominació URI (Uniform Resource Identifier = Identificador Uniforme de Recurs) per a cadascuna de les versions compilades dels recursos web digitals.

3.3.2.2. Metadades

L'assignació de metadades representa per als recursos digitals el conjunt d'aspectes relatius a la descripció catalogràfica del propi recurs (els tradicionals: títol, menció de responsabilitat, dades de publicació, descripció resum, etc.; i els específics dels nous formats: versió, arxius que conté, tipus de llenguatge de programació, etc.).

34. Muxach, S.; Lopo, A. (1999). «Metadades a peu pla». *Ítem: revista de biblioteconomia i documentació.* Núm. 24 (1999), p. 99-134. <http://www.cobdc.org/cgi-bin/intranet/itemdoc.pl?page=-num24/smuxach.pdf>. [Consulta: 25/01/2006].

35. *RECERCAT: metadades per a la descripció dels documents* (actualitzat a novembre 2005), basat en part en *Pautes i recomanacions del CBUC per a l'ús de metadades Dublín Core en recursos web* (Doc. 02/51) elaborades pel Grup de treball LIRE del CBUC.

Les metadades faciliten el camí per al que Berners-Lee denominà la *web semàntica*, o sigui, la via per aconseguir la creació d'un gran catàleg mundial de recursos on cada metadada representa normalitzadament una característica del recurs.[36]

En tot cas, i com explica Gail M. Hodge, una vegada l'arxiu ha compilat el document digital, és necessari identificar-lo i catalogar-lo.[37] La identificació proveeix una clau única que l'identifica i el relaciona amb la resta de recursos. La catalogació de les metadades dona suport a l'organització, l'accés i la conservació.

Tots els arxius —com segueix Hodge— empren algun tipus de metadada per a la seva descripció, nou ús, administració, i preservació del document arxivat. No aprofundirem aquí en els aspectes relatius a com s'ha creat la metadada, els estàndards i les normes que s'han emprat, el nivell d'aplicació de les metadades, i on estan emmagatzemades aquestes.

Sí cal deixar constància de que a partir dels models existents, la Biblioteca de Catalunya ha apostat amb la resta de membres del Consorci de Biblioteques Universitàries de Catalunya (CBUC) per definir, tenint en compte les necessitats comunes, les metadades per a la descripció de documents a partir del model Dublin Core.[38] Aquell document ha servit de base per fer la proposta de metadades del projecte, que té pendent de redefinir les especificacions en relació als materials: com en la resta de documents que formen la col·lecció de la BC, l'aplicació de les metadades és susceptible de variar depenent del tipus de recurs (pertanyent a la bibliografia nacional o no, etc.) al qual es refereixi.[39]

3.3.2.3. Emmagatzematge

El sistema preveu un dipòsit que permeti conservar tots els recursos que formen la col·lecció, de manera que s'hi pugui tenir accés en tot moment. Es contempla un sistema d'emmagatzematge per doble còpia del dipòsit a diferent ubicació geogràfica, previsiblement al CESCA i a la Biblioteca de Catalunya.

El programari emprat en el test permet avançar que els arxius s'emmagatzemen comprimits amb l'extensió estàndard.ARC.

És previst un capacitat necessària total de 10 TB en el períodes de producció i d'explotació del projecte (2006-2008).

36. Muxach, S.; Lopo, A. (1999). «Metadades a peu pla». *Ítem: revista de biblioteconomia i documentació.* Núm. 24 (1999), p. 99-134. <http://www.cobdc.org/cgi-bin/intranet/itemdoc.pl?page=num24/smuxach.pdf>. [Consulta: 25/01/2006]

37. 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: 25/01/2006]

38. *RECERCAT: metadades per a la descripció dels documents* (actualitzat a novembre 2005), basat en part en *Pautes i recomanacions del CBUC per a l'ús de metadades Dublín Core en recursos web* (Doc. 02/51) elaborades pel Grup de treball LIRE del CBUC.

39. Devem un agraïment a les aportacions i recomanacions de Francesca Navarro de la unitat de Digitalització, i de Ida Conesa del Servei de Normalització Bibliogràfica, de la Biblioteca de Catalunya.

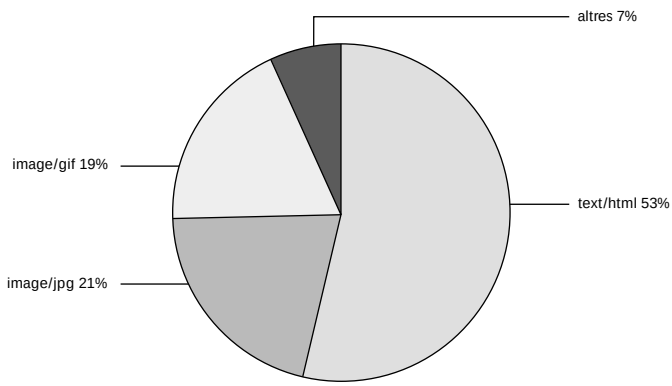
3.3.2.4. Preservació

La preservació és l'aspecte de la gestió de la col·lecció que preserva el contingut i l'aparença del document digital.

Les estratègies més habituals de preservació[40] són la migració periòdica o *refresh* de les dades (a les noves versions dels mateixos programes o llenguatges, a nous programes capaços de llegir els anteriors), l'emulació (ús del programari, especificacions, etc. utilitzat en el moment de la creació), la recreació (simulació per *enginyeria inversa* o altres mètodes). Totes aquestes tècniques tenen limitacions legals (còpia, transformació, emmagatzematge) que hauran de ser analitzades amb cura en la prova pilot del sistema.

En tot cas, les previsions sobre la tipologia d'arxius que el projecte haurà de gestionar mostren que el gruix dels arxius corresponen a formats estàndards, que poden simplificar la tasca preservadora, almenys en les macroxifres.

En un experiment realitzat per cobrir aquest detall de la comunicació, obtenim que d'una mostra d'uns 700.000 arxius (25 GB) el 96% dels arxius (17 GB) són de formats estàndards: text/html (53%), imatge jpeg o gif (21% i 19%, respectivament), o pdf (3%).



En el gràfic anterior es mostra la tipologia dels arxius capturats, i continuació es mostra el llistat dels 20 primers formats, per ordre del nombre d'arxius:

40. Ayre, C.; Muir, A. (2004). «The right to preserve: the rights issues of digital preservation». *D-Lib magazine*. Vol. 10, num. 3 (mar 2004). <http://www.dlib.org/dlib/march04/ayre/03ayre.html>. [Consulta: 25/01/2006].

URL	Gbytes	Mime-Types	URL	Gbytes	Mime-Types
357.739	4,913	text/html	1.177	0,000	text/dns
141.286	4,342	image/jpeg	788	0,956	application/zip
126.735	0,949	image/gif	653	0,342	application/octet-stream
19.777	6,707	application/pdf	650	0,018	audio/midi
3.264	1,283	text/plain	618	0,005	text/xml
3.217	0,276	application/x-shockwave-flash	552	1,813	audio/mpeg
2.942	0,023	application/x-javascript	417	0,009	no-type
2.724	0,011	text/css	374	0,001	application/xml
2.589	0,349	application/msword	348	0,019	appl./x-msdos-program
1.912	0,064	image/png	317	0,139	image/bmp

La principal dificultat, en conseqüència, estarà relacionat amb els formats o aplicacions que són efímers,[41] conscientment o no. Podem avançar que és en fase de desenvolupament la captura i preservació de certes pàgines web dinàmiques (java script), amb imatges en moviment (streamed video) o so (streamed audio), xat, conferències en xarxa, subhastes en línia, videojocs, comerç electrònic, etc.

El sistema contemplarà els aspectes relatius a la preservació dels recursos digitals, tenint en compte que els principals perills en els formats digitals estan relacionats amb la pèrdua de la capacitat de veure *tal com es va crear* el producte digital (*look & feel*). Per citar alguns dels obstacles:

— Envelliment, obsolescència o degradació de les versions del programari o llenguatge utilitzat en la producció dels continguts
— Ús de software propietari
— Pèrdua del *context* o links on s'emmarca el recurs
— Bases de dades dinàmiques
— Accessos per *cookies*, contrasenya o control d'IP, etc.

3.3.3. Accés permanent als recursos

L'accés als recursos del projecte, en línia i en obert, està limitada a allò que es recomani en els serveis jurídics a tal efecte. En tot cas, és previst arribar a acords amb un nombre indeterminat d'editors de la web per assegurar-ne aquest accés que, en un servei

41. A Christensen, N. (2005). «Preserving the bits of the Danish Internet». *Internet Web Archiving Workshop* (5th: 2005: Viena). <http://www.iwaw.net/05/christensen.pdf>. [Consulta: 25/01/2006], s'ha desenvolupat el concepte de «Vida útil del dipòsit» (MTTF, *Mean Time to Failure*), aplicat a projectes com el que ens ocupa, per calcular el temps de vida útil abans que es produeixi una falla irreversible que comenci a deteriorar parts del dipòsit. En el cas danès aquest període s'estipula actualment en 144 anys.

de mínims i sense els preceptius acords, es podria realitzar a les dependències de la Biblioteca de Catalunya.

La recuperació de la informació està assegurada per la catalogació per metadades, que pot possibilitar la integració dels recursos al catàleg bibliogràfic de la BC, i per la capacitat de cerca lliure en el text dels recursos que possibilita l'indexador del paquet de programari que previsiblement s'emprarà en el sistema.

3.4. Fases d'implementació

El 2005 la Biblioteca de Catalunya va iniciar la fase preliminar, de planificació, en la qual s'ha realitzat l'anàlisi dels projectes i recursos existents, els agents implicats en la producció de pàgines web a Catalunya, i els aspectes legals que en condicionen les pràctiques que es volen dur a terme. Paral·lelament, s'ha realitzat una sèrie de proves de maquinari i programari, que permeten a la Biblioteca de Catalunya detectar quina és la seva capacitat en matèria de captura i procés de les dades. Finalment, s'ha documentat l'escenari previst de funcionament del sistema, així com els recursos econòmics, tecnològics i de personal adients.

L'any 2006 representa la fase de producció, en la qual és previst realitzar el pla pilot del projecte, amb la plena integració del soci tecnològic, el CESCA, al PADICAT. En aquest sentit, l'objectiu és realitzar diverses accions de captura exhaustiva de la web catalana, així com arribar als primers acords de cooperació amb els agents públics i privats, susceptibles de gestionar el dipòsit voluntari regular de les seves pàgines web. Addicionalment, s'estudia la possibilitat de focalitzar el projecte al voltant d'un determinat esdeveniment d'interès públic.

Els anys 2007 i 2008 s'inclouen en la fase d'explotació i creació de l'oficina PADICAT, per tal de realitzar sistemàticament la captura de la web catalana, així com ampliar els acords amb els productors i tanmateix les accions dirigides a gestionar exhaustivament determinats esdeveniments (socials, culturals, polítics, etc.) susceptibles de crear línies de recerca futures.

El 2009 ha de permetre la Biblioteca de Catalunya i els seus socis de projecte comptar amb un escenari òptim, en el qual funcioni a ple rendiment aquest sistema, que haurà estat pioner a Espanya i tanmateix de referència a Europa,[42] amb uns indicadors quantitatius de 100.000 pàgines web capturades en diverses edicions, que possiblement signifiquin uns 50 milions d'arxius i 30 Terabytes de volum. Són objectius paral·lels el tancament d'acords de cooperació amb 300 institucions de tot tipus, així com permetre l'accés en obert, en línia, a bona part de la col·lecció.

42. La web *territorial*, aplicable a llengües transeuropees, territoris infraestatals o transestatals, temàtiques concretes, dominis no geogràfics, etc.

4. Beneficis del projecte

Amb l'escenari descrit, els beneficis a ulls dels lectors bibliotecaris-documentalistes de Catalunya, els professionals de la informació i la documentació, són evidents: confecció de la bibliografia nacional més enllà dels formats tradicionals i 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 al sistema bibliotecari, possibilitats infinites de cooperació amb la resta de biblioteques, arxius i museus de Catalunya; impuls i lideratge en la confecció del patrimoni digital d'Espanya. Finalment, 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 són fruit del coneixement i l'expressió dels creadors del segle XXI, ja siguin de caràcter cultural, educatiu, científic o administratiu, o comprenguin informació tècnica, jurídica, mèdica i d'un altre tipus.

Bibliografia

AGENJO, X.; HERNÁNDEZ CARRASCAL, F. (2005). «La recolección de metadatos (metadata harversting) y su aplicación en España». En: Jornadas Españolas de Documentación. (9es: 2005: Madrid). *Fesabid 2005: Infogestión.* [Madrid]: Fesabid, 2005, p. 237-254.

AYRE, C.; MUIR, A. (2004). «The right to preserve [en línia]: the rights issues of digital preservation». *D-Lib magazine*, vol. 10, no. 3 (March 2004). <http://www.dlib.org/dlib/march04/ayre/03ayre.html>. [Consulta: 25/01/2006].

BAEZA-YATES, R.; CASTILLO, C.; LÓPEZ, V. (2005). «Characteristics of the Web of Spain» [en línia]. *Cybermetrics*, vol. 9, issue 1, paper 3 (2005). <http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html>. [Consulta: 25/01/2006].

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

BOYKO, A. [et al.] (2004). *Test bed taxonomy for crawler* [en línia]. Version 1.0. International Internet Preservation Consortium, July 2004. 15 p. <http://netpreserve.org/publications/iipc-r-002.pdf>. [Consulta: 25/01/2006].

CHRISTENSEN, N. (2005). «Preserving the bits of the Danish Internet» [en línia]. En: Internet Web Archiving Workshop (5è: 2005: Viena). *5th International Web Archiving Workshop*. <http://-www.iwaw.net/05/christensen.pdf>. [Consulta: 25/01/2006].

CHRISTENSEN-DALSGAARD, B.; Fønss-Jørgensen, Eva; Hielmcrone, Harald von [et al.] (2003). *Final report for the pilot project «netarkivet.dk»* [en línia]. [Kobenhavn: The Royal Library], Feb. 2003. [63] p. <http://netarkivet.dk/publikationer/webark-final-rapport-2003.pdf>. [Consulta: 25/01/2006].

CORDÓN, J.A. (2005). «El depósito legal y los recursos digitales en línea» [en línia]. En: Jornadas sobre Bibliotecas Nacionales (2005: València). *Las bibliotecas nacionales del siglo XXI*. Valencia: Biblioteca Valenciana, 2005. 20 p. <http://bv.gva.es/documentos/Ponencias/Cordon.pdf>. [Consulta: 25/01/2006].

Directrices para la preservación del patrimonio digital [en línia] (2003). Canberra: Unesco, 2003. 186 p. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 25/01/2006].

DULABAHN, B. (2004). «The National Digital Information Infrastructure and Preservation Program (NDIIP) [en línia]: future directions and relevance to other countries». En: *Archiving web resources*. Canberra: National Library of Australia. <http://www.nla.gov.au/webarchiving/>. [Consulta: 25/01/2006].

DUPONT, H. (1999). «Legal diposit in Denmark [en línia]: the new law and electronic products». *LIBER quarterly, the journal of the European research libraries*, vol. 9, no. 2, (1999). <http://libermaps.kb.nl/articles/dupont11.htm>. [Consulta: 25/01/2006].

GOMES, D.; SILVA, M.J. (2005). «Characterizing a national community Web» [en línia]. *ACM transactions on Internet technology*, vol. 5, no. 3 (August 2005). <http://xldb.fc.ul.pt/daniel/gomes-Characterizing.pdf>. [Consulta: 25/01/2006].

HAETTIGUER, M. (2003). «Vers la conservation des sites web régionaux» [en línia]. *BBF*, t. 48, no. 4 (2003), p. 77-84. <http://www.enssib.fr/bbf/bbf-2003-4/13-haettiger.pdf>. [Consulta: 25/01/2006].

HALLGRÍMSSON, P.; BANG, S. (2003). «Nordic Web Archive» [en línia]. En: *International* Web Archiving Workshop (3r: 2003: Trondheim). *3rd ECDL Workshop on Web Archives.* <http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf>. [Consulta: 25/01/2006].

HODGE, G.M. (2000). «Best practices for digital archiving [en línia]: an information life cycle approach». *D-LIB Magazine,* vol. 6, no. 1 (Jan. 2000). <http://www.dlib.org/dlib/january00/01hodge.html>. [Consulta: 25/01/2006].

KIMPTON, M. (2004). «Saving the web for future generations» [en línia]. En: *Archiving web resources*. Canberra: National Library of Australia. <http://www.nla.gov.au/webarchiving/>. [Consulta: 25/01/2006].

LLUECA, C. (2005). «Webs sempre accessibles [en línia]: 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].

MANNERHEIM, J. (2004). «Collect all, catalogue some» [en línia]. *Archiving web resources*. Canberra: National Library of Australia. <http://www.nla.gov.au/webarchiving/>. [Consulta: 25/01/2006].

MARILL, J. [et al.] (2004). *Web harvesting survey* [en línia]. Version 1. International Internet Preservation Consortium, July 2004. 10 p. <http://netpreserve.org/publications/iipc-r-001.pdf> [Consulta: 25/01/2006].

MASANÈS, J. (2004). «International Internet Preservation Consortium (IIPC) [en línia]: web archiving toolset». En: *Archiving web resources*. Canberra: National Library of Australia. <http://www.-nla.gov.au/webarchiving/MasanesJulien.ppt>. [Consulta: 25/01/2006].

MASTERS, R. [et al.] (2005). «The large-scale archival storage of digital objects» [en línia]. En: *Digital Preservation Coalition Meeting.* (2005: York). <http://www.dpconline.org/graphics/events/-050422meeting.html>. [Consulta: 25/01/2006].

MOHR, G. [et al.] (2004). «An introduction to Heritrix [en línia]: an open source archival quality web crawler». En: *International Web Archiving Workshop* (4t: 2004: Bath). <http://www.iwaw.net/04/proceedings.php?f=Mohr>. [Consulta: 25/01/2006].

MUXACH, S.; LOPO, A. (1999). «Metadades a peu pla» [en línia]. *Item: revista de biblioteconomia i documentació.* Núm. 24 (gener-juny 1999), p. 99-134. <http://www.cobdc.org/cgi-bin/intranet/itemdoc.pl?page=num24/smuxach.pdf>. [Consulta: 25/01/2006].

PAREJA, V.M. [et al.] (2005). «Desarrollo y aplicación del concepto de sede web como unidad documental de análisis en Cibermetría». En: Jornadas Españolas de Documentación (9es: 2005: Madrid). *Fesabid 2005: Infogestión.* [Madrid]: Fesabid, 2005, p. 325-338.

PHILLIPS, M. (2004). «What to collect and how to do it [en línia]: the National Library of Australia's selective approach». En: *Archiving web resources*. Canberra: National Library of Australia. <http://www.nla.gov.au/webarchiving/>. [Consulta: 25/01/2006].