Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)

Biblioteca de Catalunya, desembre de 2005

Memòria de plantejament del projecte PADICAT

Sumari

1. Introducció a la Memòria de plantejament del projecte PADICAT

El present informe és el resultat d'una sèrie de treballs realitzats en relació a l'adjudicació del servei de consultoria i assistència tècnica per al disseny del projecte PADICAT (Patrimoni Digital de Catalunya), que la Biblioteca de Catalunya cobria el juny de 2005 amb l'expedient G0883 N07/05.

El servei incloïa en les activitats vinculades al projecte la presentació d'un seguit de treballs, a partir dels quals s'ha creat el present informe global.

Concretament, i com es recull al sumari precedent, l'informe conté:

- Estat de la qüestió arreu del món
- Context a Catalunya: recursos existents, agents implicats, aspectes legals
- Disseny del sistema d'informació: abast, captura, organització i accés als recursos
- Maquinari i programari necessari: plataformes i opcions (inclòs testeig de programari)
- Planificació a curt i mig termini i fases d'execució
- Recursos humans necessaris per executar el projecte: perfils i tasques
- Estudi de costos vinculats a la fase de producció

És objectiu del present informe presentar de manera unificada les qüestions anteriors, tot possibilitant un punt de partida per a la fase de producció del projecte, prevista per a l'any 2006.

2. Estat de la qüestió arreu del món[1]

2.1. Introducció

Els nombrosos avantatges que faciliten les tecnologies d'informació i comunicació (TIC) han possibilitat que el patrimoni cultural, científic i d'informació es presenti en format digital, en detriment dels formats analògics tradicionals.

Un repositori nacional té la missió de garantir l'accés a llarg termini als recursos digitals que es generen a un territori, o sobre un territori determinat.

Cada vegada més, els recursos fruit del coneixement o l'expressió dels éssers humans, siguin aquests de caràcter cultural, educatiu, científic o administratiu, o englobin informació tècnica, jurídica, mèdica i d'altres classes, es generen directament en format digital o es converteixen en aquest a partir del material analògic ja existent.

Els productes "de naturalesa digital" no existeixen en un altre format que no sigui l'electrònic original.[2]

Aquesta realitat, sumada a la voluntat de les persones, les institucions i els governs de vetllar per la preservació de qualsevol forma de patrimoni, ha facilitat que les administracions de diversos països hagin endegat polítiques destinades a garantir l'accés permanent (la recopilació i emmagatzematge, el tractament, i la preservació) a la producció digital, per part dels agents públics i privats.

Les dificultats són notables: per començar, els mètodes tradicionals de preservació de la producció bibliogràfica (com el dipòsit legal) són de difícil aplicació perquè a banda de la possible obsolescència del text legal (el cas espanyol), els recursos digitals poden instal·lar-se en servidors d'arreu del món, el que dificulta la tradicional correspondència geogràfica entre la ubicació del productor i la llengua o la temàtica publicada. En segon lloc, el volum de la producció digital té un volum i un creixement exponencial, essent a més molt variable la durabilitat dels materials publicats a Internet[3], i en conseqüència, limitada la possibilitat d'accés permanent al patrimoni. Finalment, assenyalar la qüestió de la propietat intel·lectual del producte digital, mancat d'un dret basat en el principi de còpia per a preservació que possibiliti la preservació del patrimoni digital, amb les limitacions comercials que siguin necessàries.

---

[1] A partir del present capítol de l'informe PADICAT es va publicar l'article: Llueca Fonollosa, Ciro (2005). «Webs sempre accessibles : les biblioteques nacionals i els dipòsits digitals nacionals». *BiD: textos universitaris de biblioteconomia i documentació*, desembre, núm. 15. <http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec1.htm> [Consulta: 01-02-2006].

[2] Traducció lliure de les *Directrices para la preservación del patrimonio digital* [en línia]. Canberra: Unesco, marzo 2003. [consulta juny 2005:] http://unesdoc.unesco.org/images/0013/001300/130071s.pdf

[3] En 44 dies fixa la vida d'una pàgina web l'UK Web Archiving Consortium: http://info.webarchive.org.uk/pressrelease21-06-04.html [consulta juny 2005]

Com s'ha apuntat, i malgrat les dificultats, diversos països han entès la necessitat de *passar a l'acció*, d'emprendre polítiques i accions de preservació per assegurar la pervivència de la producció digital, com ja s'havia fet històricament amb els documents impresos i en suports tradicionals, amb les lleis nacionals del dipòsit legal. En la major part dels casos que es presentaran ha estat la *biblioteca nacional* qui ha liderat el procés de preservació i accés del patrimoni digital, implicant la resta d'agents.

En els orígens de la preservació digital podem esmentar les accions de les *biblioteques virtuals*, aquells projectes de les biblioteques (de recerca, universitàries, nacionals, i també públiques) dedicats a presentar directoris temàtics de recursos electrònics. Complementàriament, es va optar per crear dipòsits multiformat (imatges, so, text, gràfics, etc.) per donar resposta a necessitats temàtiques concretes, normalment d'àmbit geogràfic, seguint el model enciclopèdic digital dels Cd-roms. El pas lògic següent ha estat conservar els recursos propis per garantir-ne l'accés amb totes les variables formals que s'han produït en el temps, en el que es coneix com els repositoris institucionals (l'arxiu web de la pròpia institució, habitualment centres universitaris que contenen la producció científica dels propis professors i centres de recerca). Quan el procés es dedica a un territori, parlem dels repositoris nacionals, arxius web, o biblioteques nacionals digitals.

Un dipòsit nacional, així, té la missió de garantir l'accés a llarg termini als recursos digitals que es generen a un territori, o sobre un territori determinat. I de fet, la missió de la Biblioteca de Catalunya (BC) no és altra que *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*.

Aquest *patrimoni bibliogràfic* que incorpora la missió de la BC inclou també la *producció bibliogràfica digital catalana*, que conformarà el PADICAT, el Patrimoni Digital de Catalunya.

2.2. Els models existents

Les experiències existents, si ens centrem en els repositoris nacionals, s'agrupen en dos models inicials: l'integral o exhaustiu (model majoritari, i característic de Suècia, Noruega, Finlàndia, Islàndia i Àustria, entre d'altres), que aposta per la integració automàtica del *total* de la web objecte de preservació, en base a determinats criteris infraestructurals (lingüístics, segons el domini de les web, segons ubicació del servidor, etc.); i el selectiu (assimilat per Austràlia, Canadà, Japó i el Regne Unit, entre d'altres), que dirigeix les accions de recopilació en base a una política selectiva temàtica (sobre un espai geogràfic determinat, al voltant d'un tema d'interès nacional, etc.), i arribant a acords amb els editors web.

Dos models inicials de la preservació web: l'integral, que aposta per la integració automàtica total de la web, i el selectiu, que dirigeix les accions en base a una política selectiva temàtica, han deixat pas a un model híbrid, que conjuga els anteriors.

El model integral

Les principals avantatges o punts forts del model integral són:

- Riquesa de la col·lecció, en quantitat i en qualitat, atès que no es condiciona selectivament que *és interessant* i que no ho és, màxim considerant que difícilment som capaços de preveure ara quins seran els usos i les línies d'investigació futures, o sigui el potencial valor de la futura recerca. En reflectir, a més, el creixement de la web i els canvis en el disseny de la publicació web de tots nivells, aporta un component sociològic afegit, ja que les pàgines personals, els *weblogs*, els xats, i *in extremis* els videojocs en línia, formen part també de la producció digital nacional en els *repositoris integrals*. Lligat a aquest punt fort hi ha el fet que la captura exhaustiva permet respectar la interrelació de les seus web.

El projecte suec, Kulturarw3, inclou les seus web amb domini .se (Suècia), .nu ("ara", en suec i altres llengües escandinaves, molt utilitzat), les seus web amb dominis internacionals (.com, .org, .net) ubicades a servidors al territori[4] *suec, i la* Suecana extreana*: les webs que parlen sobre Suècia, viatges per Suècia, o traduccions d'obres literàries sueques.*

- Compilació automàtica. El maquinari i programari emprat assegura una capacitat alta de captura i emmagatzematge en resposta a uns paràmetres determinats, que poden ser tan amplis com la direcció del projecte estableixi. Aquest fet minimitza els recursos més costosos, els personals, alhora que aconsegueix resultats visibles (resultats presentables, vendibles a la sensibilitat política i ciutadana) en un període de temps raonablement curt: es crea en un temps raonable una col·lecció digital.

Finalitzada la primera captura del projecte finlandès (agost-setembre de 2001), els responsables de la Biblioteca Nacional de Finlàndia asseguren que s'han capturat 7.5 milions d'URL, i calculen que aquesta fotografia *de la web finlandesa representa entre el 30 i el 50% del total capturable.*

- Baix cost. En relació amb els aspectes ja descrits anteriorment, s'observa que els projectes que aposten per un model integral estan coordinats per equips petits de persones. Malgrat que aquests equips estan integrats a les estructures de les biblioteques nacionals, el cert és que no dediquen els recursos personals que es requereixen per a la catalogació i la gestió dels processos selectius.

Mentre que el projecte australià PANDORA (selectiu) ocupa un total de 13 persones a temps complet, i el sistema emprat a Quebec (selectiu, i integrat al catàleg IRIS) presenta un equip de 10 persones, els

[4] Eines com Maxmind (www.maxmind.com) o Ip2location (www.ip2location.com) donen la ubicació geogràfica dels servidors [consulta setembre 2005]. Les webs amb domini .se són el 53% de l'arxiu. Les .nu representen el 17%, la resta, el 40%.

projectes basats en la captura integral presenten equips sensiblement menors: els casos coneguts d'Àustria, i Suècia, amb equips d'1 a 3 persones.

Els principals inconvenients o punts febles del model integral són:

- Impossibilitat d'accedir a la Internet invisible, doncs el sistema automàtic només accedeix als recursos que són publicats en obert, i per tant es produeixen llacunes en la captura de webs de pagament, protegides amb contrasenyes, pàgines orfes, la major part de les dinàmiques; així com la impossibilitat d'accedir a la Infranet: les bases de dades bibliogràfiques (catàlegs de biblioteques) o alfanumèriques (diccionaris).

*Com ha explicat repetidament Isidro Aguillo i el personal de l'*InternetLab[5]*, la Internet invisible multiplica la Internet visible, segons els autors i estudis, entre 2 i 50 vegades, i el més important: suposa un arxipèlag de qualitat, pel tipus de recursos que s'hi contenen: articles, estudis científics, publicacions digitals, etc.*

- Compilació irregular de la col·lecció, ateses les llacunes en el control dels ítems col·lectats (per exemple, en les publicacions periòdiques), la no reclamació dels documents no accessibles, i la pèrdua de documents importants i dels canvis freqüents que es produeixen a determinades seus web.

El projecte noruec Paradigma va començar les seves captures integrals el 2001 amb el domini .no. En la tercera captura (agost 2003) s'ha ampliat l'abast als dominis internacionals (.com, .net) i a 65 diaris digitals noruecs que quedaven exclosos de les captures periòdiques. És un exemple d'un model integral amb accions selectives.

- Accés limitat als resultats. D'una banda, la manca d'un procés de catalogació pel sistema que sigui (metadades a diferents nivells, inclusió al catàleg de la biblioteca, etc.), dificulta la recuperació dels documents capturats. D'altra, el respecte als drets d'autor per publicar sense autorització (sense acord previ, en tot cas) els recursos capturats fa que es restringeixi l'accés als dipòsits nacionals, normalment a les pròpies instal·lacions de les biblioteques nacionals: aquesta mesura difícilment conjuga amb la pròpia naturalesa dels projectes: garantir l'accés a la producció digital.

Únicament Internet Archive *ofereix accés obert, en línia, als seus fons, i només permet la cerca per URL. La resta de projectes, inclosos els més veterans –els escandinaus— limiten la consulta de les seves col·leccions a les dependències de les biblioteques nacionals que els lideren. Aparentment,*

[5] InternetLab pertany al CINDOC: http://internetlab.cindoc.csic.es/ [consulta juny 2005]

només una part molt petita de la col·lecció rep un tractament que possibiliti la recuperació més enllà de l'URL o la data de captura.

El model selectiu

Les principals avantatges o punts forts del model selectiu són:

- Creació d'una col·lecció equilibrada, atès que cada ítem que formarà part de l'arxiu és avaluat tenint en compte el creixement de la col·lecció. Respon per tant a un model més clàssic bibliotecari, en el sentit que coneix el que forma part del seus fons i l'amplia tenint en compte la realitat del territori i tots els usuaris.

El model de creixement de PANDORA és visible també en UKWA (Regne Unit), atès que proposa una classificació dels continguts molt similar al popular directori Yahoo. A partir d'un màxim de 9 categories inicials (Art i humanitats, negocis i economia, etc.), es produeix una taxonomia en cascada (Art i humanitats: Arquitectura; Dansa; Belles Arts; Geografia; Història; Llengües; Literatura; Música...) i la presència dels recursos digitals britànics és temàticament equilibrada en totes elles.

- Màxima facilitat d'accés al fons, doncs cada ítem por ser completament catalogat i formar part de la bibliografia nacional: les dades bibliogràfiques poden ser compartides. Al propi catàleg de la biblioteca els recursos estan integrats, i els acords permeten publicar els recursos en línia, obertament. En tot cas, la catalogació dels documents fa que les possibilitats de recuperació siguin il·limitades.

El juny de 2004, el projecte WARP (Japó) ofereix accés complet a 600 seus web (administració, universitats, congressos i seminaris) i 110 diaris electrònics. Lituània té el seu arxiu de recursos electrònics integrat en el catàleg col·lectiu LIBIS. Finalment, PANDORA (Austràlia), té integrat el seu fons al catàleg de la biblioteca, i permet als cercadors (Google, Msn, etc.) accedir a determinats nivells dels recursos.

- Estratègic. En funcionar amb aliances i acords amb les entitats editores (comercials o no) la implicació dels agents productors es produeix més naturalment, amb voluntat compartida. A banda, els acords fan possible que els ítems siguin accessibles en línia en tota la seva profunditat, alhora que la informació formal i propietats de cada recurs són coneguts pels gestors de captura, el que permet desenvolupar mètodes i eines de compilació i accés, així com les estratègies de preservació a més llarg termini. La web invisible i la Infranet s'hi inclouen al dipòsit.

Possiblement sigui PANDORA (Austràlia) l'exemple més evident, que no l'únic, de la força de la cooperació. La Biblioteca Nacional d'Austràlia compta amb diverses institucions sòcies de projecte:

Australian War Memorial, Australian Institute of Aboriginal and Torres Strait Islander Studies, i les biblioteques dels diferents estats del país, entre d'altres. L'aliança proporciona rigor en la selecció, suport als pressupostos, alhora que presència mediàtica i en la comunitat de la recerca australiana

Els principals inconvenients o punts febles del model selectiu són:

- Parcialitat en descriure el món: en la selecció dels recursos que es recopilaran, es realitza un judici subjectiu sobre el valor dels recursos i el que els investigadors preferiran en el futur. En tot cas, l'extensió d'un arxiu selectiu és molt limitat en comparació amb el volum del material d'un territori determinat, i malgrat els esforços, els criteris de selecció són de difícil definició.

El projecte britànic UK Web Archive està liderat per la British Library i compta, entre d'altres importants socis, amb la National Library of Wales i la National Library of Scotland. El projecte està en desenvolupament, i ha estat fet públic molt recentment (maig 2005), però el fet és que alguns ítems seleccionats pels socis de projecte donen una visió parcial de la web britànica: teatre (únic ítem disponible: Theatre in Wales).

- Elevat cost, atès que la selecció, la gestió i el seguiment dels acords i captures, i especialment l'anàlisi documental dels recursos és una tasca molt intensiva i el cost per ítem és elevat en recursos humans. El fet que les institucions que gestionen els repositoris siguin les biblioteques nacionals garanteixen una alta qualitat en la descripció i indexació dels recursos, habitualment per llenguatge de metadades.

En el congrés celebrat a Canberra[6] el novembre de 2004, la responsable de l'australià PANDORA desvetllava que el cost per a la gestió d'un ítem digital pot arribar a ser cinc vegades superior que el d'una monografia.

- Descontextualització de la col·lecció, perquè la selecció dels recursos no necessàriament es realitza en el seu context i per tant no inclou els recursos *linkats* que contextualitzen la informació. En el llenguatge de l'hipertext, la selecció d'una web sense tenir en compte amb quines altres està lligada pot donar una lectura *orfe* del recurs.

El principal problema de l'emblemàtic PANDORA és la preservació dels links trencats. La política que se segueix és donar la possibilitat a l'usuari d'accedir a la versió actual de la web a la qual apuntava el link del recurs preservat, però els problemes de contextualització no queden resolts.

El model híbrid

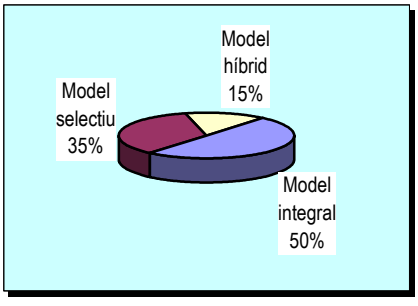
Aquests dos models anteriors han deixat pas en alguns països a models híbrids, (França, Dinamarca, Nova Zelanda) que complementen la presa periòdica del total de la web nacional amb acords segons interessos temàtics, i tanmateix en relació a qüestions d'actualitat (eleccions, catàstrofes, etc.).

Model híbrid: el cas de Dinamarca	Captura exhaustiva de la web danesa (*.dk) + Acords amb entitats editores (administració, universitats, etc.) + Captura integral d'events específics (eleccions, esports, etc.)

De l'estudi detallat dels repositoris existents es desprèn que aquesta és la tendència a seguir per la majoria dels projectes integrals (Àustria, Països baixos, Suècia, Finlàndia, etc.) Lògicament, els projectes híbrids equilibren alguns dels avantatges descrits anteriorment (col·lecció rica i equilibrada, màxim accés, impuls dels acords estratègics, compilació automatitzada i seguiment de les llacunes), però també complica o no supera elements superats (cost elevat, equips més nombrosos, càrrega de gestió).

Finalment, altres anàlisis teòrics[7] de la situació apunten a una classificació més complexa (segons si la web a capturar és estàtica o estàtica, per exemple), però entenem aquí que la captura restrictiva de la web segons la seva complexitat (si és estàtica, és més fàcil capturar-la i preservar-la) de preservació i garantia d'accés és només un primer pas per al destí posterior de tots els repositoris nacionals.

2.3. Dipòsits digitals nacionals



S'ha trobat referències de vint casos de repositoris nacionals: Alemanya, Austràlia, Àustria, Canadà, Dinamarca, Estats Units d'Amèrica, Estònia, Finlàndia, França, Grècia, Islàndia, Japó, Lituània, Noruega, Nova Zelanda, Països baixos, Quebec, Regne Unit, República Txeca, i Suècia:

- De model integral: Alemanya, Àustria, Estònia, Finlàndia, Grècia, Islàndia, Lituània, Noruega, República Txeca, i Suècia.

[6] *Archiving web resources* (Canberra : nov 2004) http://www.nla.gov.au/webarchiving/ [consulta juny 2005]
[7] Cordón, José Antonio. "El depósito legal y los recursos digitales en línea", *Las Bibliotecas Nacionales del siglo XXI*. [en línia]. Valencia: Biblioteca valenciana, 2005. [consulta juny 2005:] http://bv.gva.es/documentos/Ponencias/Cordon.pdf

- De model selectiu: Austràlia, Canadà, Estats Units d'Amèrica, Japó, Països baixos, Quebec i Regne Unit.

- Finalment, considerem models plenament híbrids els repositoris de Dinamarca, França, i Nova Zelanda.

Com s'ha esmentat però, la major part dels repositoris que considerem segueixen un model integral han adoptat mesures per incloure determinats recursos (publicacions periòdiques) que els fan acostar-se a paràmetres híbrids. És la tendència generalitzada.

És precís esmentar dos recursos que proporcionen informació detallada de cadascun d'aquests projectes:

- El portal PADI[8] de la National Library of Australia. El recurs conté informació actualitzada de la pràctica totalitat dels projectes existents, així com informació diversa dels aspectes relacionats amb la preservació web (dipòsit legal, eines, bibliografia, etc.). A banda, la Biblioteca Nacional d'Austràlia va organitzar el novembre de 2004 el congrés *Archiving web resources*[9] en el qual és accessible en obert la major part de presentacions realitzades en aquella activitat.ç

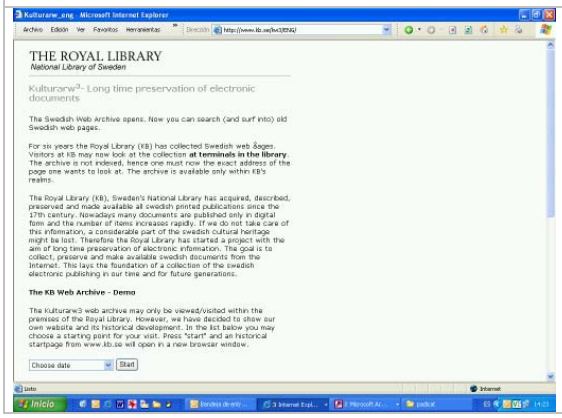
- IWAC[10] (International Web Archiving Workshop) se celebra anualment, i està organitzat per un grup de professionals procedents dels diversos projectes. En les edicions anteriors a Viena (setembre 2005) és consultable la documentació relativa a la presentació dels projectes que existeixen.

Analitzarem a continuació els casos existents, centrant-nos en els tres exemples que representen els models anunciats: integral (Kulturarw3 de Suècia), selectiu (PANDORA d'Austràlia), i híbrid (Netarchive.dk de Dinamarca); i citant la resta de projectes amb una descripció somera de les seves característiques. No s'aporten dades d'Islàndia ni d'Estònia per manca de bibliografia.

[8] *PADI: Preserving Access to Digital Information*. http://www.nla.gov.au/padi/ [consulta juny 2005]
[9] *Archiving web resources* (Canberra : nov 2004) http://www.nla.gov.au/webarchiving/ [consulta juny 2005]
[10] *International web archiving workshop*. http://www.iwaw.net [consulta juny 2005]

Model integral

Kulturarw3



Inici	Suècia, 1996
Lidera	Kung.Royalbiblioteket
Contacte	Allan Arvidsson (allan.arvidson@kb.se)
Model	Integral
Descripció	Exhaustiu en la captura de la web sueca: 350.000 webs (feb 2005). L'accés al fons està limitat a les dependències de la Biblioteca Nacional de Suècia. La catalogació dels materials no és una prioritat.
URL	http://www.kb.se/kw3/ENG/

El cas suec és paradigma d'anticipació. A partir dels orígens del dipòsit legal, de 1661, la revisió de 1993 inclou la informació electrònica publicada en suports (fitxer informàtic, CD-ROM). La Biblioteca nacional sueca crea el 1996 el Kulturarw3, l'Arxiu Web Suec. Sis anys més tard, el 2002, es decreta a Suècia que la biblioteca nacional realitza *de iure* els treballs de preservació i accessibilitat permanent del patrimoni digital suec.

La col·lecció cobreix revistes digitals i publicacions periòdiques no diàries, així com, des de fa uns mesos, una selecció de més de 100 títols de diaris suecs, documents estàtics (arxius electrònics), i documents dinàmics amb links. Ulteriorment es recopila el contingut de llistes de discussió, i arxius *ftp oberts*.

Les eines de captura i organització són el programari Combine i, més recentment per als diaris, Heritrix.

Per les dades fetes públiques (febrer 2005), sabem que el darrer volum del Kulturarw3 és de 306 milions d'arxius i uns 10.000 Gb: 350.000 seus web. Consultable únicament a les dependències de la Biblioteca nacional sueca.

Els punts forts[11] del projecte Kulturarw3 són els derivats de l'exhaustivitat en la compilació automàtica i el valor en plasmar la societat *digital* sueca. De les revistes científiques als *weblogs* de les ONG. Els punts febles estan relacionats amb importants llacunes en el control del que es captura, la nul·la profunditat en la *Infranet* (continguts de pagament o amb contrasenya, pàgines orfes, etc.) i la manca

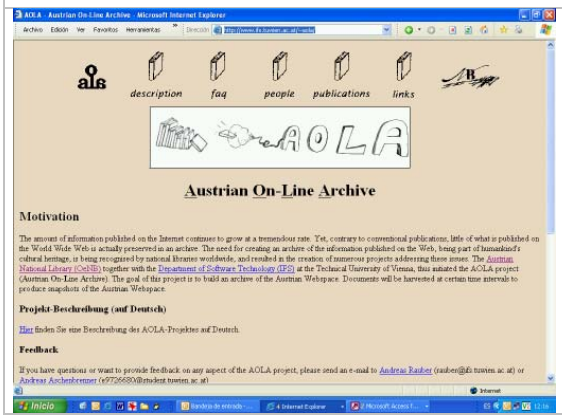
[11] Mannerheim, Joan. "Collect all, catalogue some", *Archiving web resources: international conference* (Canberra: nov 2004). [en línia]. Canberra: National Library of Australia, 2005. [consulta abril 2005:] http://www.nla.gov.au/webarchiving/program.htm

de catalogació de l'arxiu. Com s'ha apuntat, el fet que l'arxiu sigui només consultable a les dependències de la biblioteca sueca és una característica negativa del sistema suec.

DEPOSIT.DDB.DE

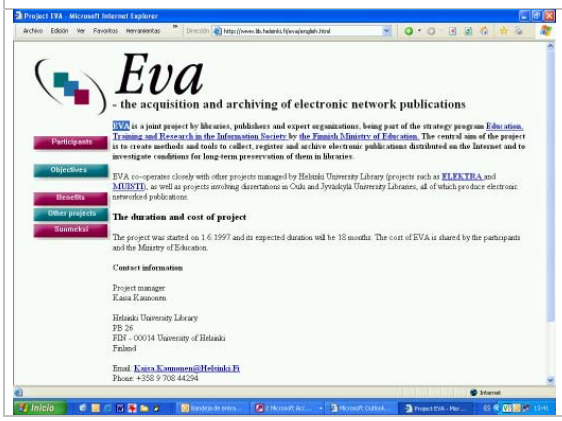
Inici	Alemanya, 1997
Lidera	Die Deutsche Bibliothek
Contacte	Hans Liegmann (liegmann@dbf.ddb.de)
Model	Integral a híbrid
Descripció	Les proves inicials es van realitzar amb la web del Govern alemany. A partir de 2002 s'arriba a acords amb editors alemanys. Catalogació per metadades.
URL	http://deposit.ddb.de/online/vdr/titel.htm

AOLA



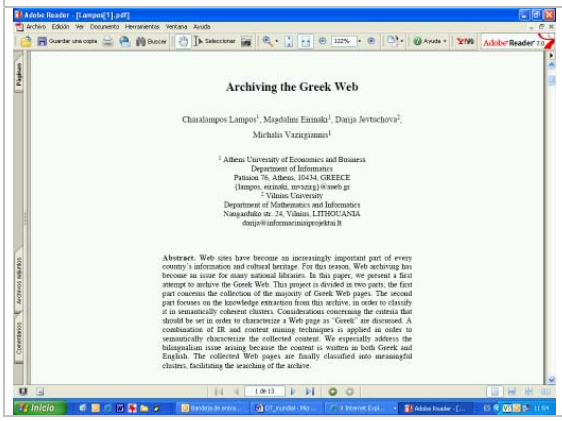
Inici	Àustria, 1999
Lidera	Österreichische National-bibliothek
Contacte	Andreas Rauber (rauber@ifs.tuwien.ac.at)
Model	Integral a híbrid
Descripció	El projecte ha patit aturades per manca de fons. El programari NEDLIB de la primera fase va donar lloc al COMBINE en una etapa posterior. El creixement previst és de 7 Gb diaris.
URL	http://www.ifs.tuwien.ac.at/~aola/

EVA



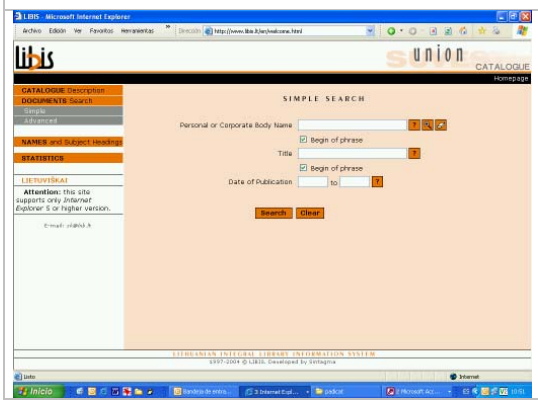
Inici	Finlàndia, 1997
Lidera	Helsingin yliopiston kirjasto (Kansalliskirjasto)
Contacte	Juha Hakala (juha.hakala@helsinki.fi)
Model	Integral a híbrid
Descripció	El projecte EVA (1997), adreçat en successives etapes a publicacions periòdiques, va donar pas el 2001 a l'arxiu web, que inclou el domini *fi. La Biblioteca Nacional de Finlàndia lidera el NWI (Nordic Web Index), que pretén ser l'arxiu web escandinau.
URL	http://www.lib.helsinki.fi/tietolinja/0203/webarchive.html

Greek Web



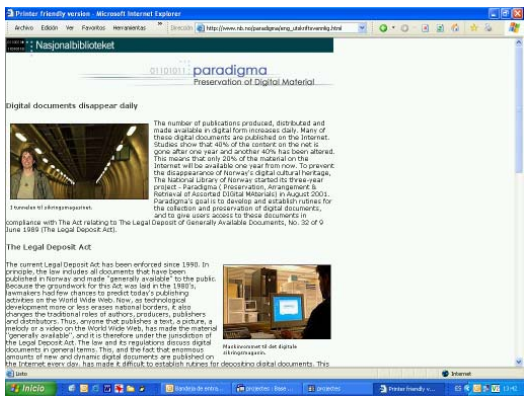
Inici	Grècia, 2003
Lidera	Athens University of Economics and Business
Contacte	Charalampos Lampos (lampos@aueb.gr)
Model	Integral
Descripció	El projecte grec és un experiment destinat a capturar el domini grec, amb programari propi.
URL	http://www.iwaw.net/04/proceedings.php?f=Lampos

LIBIS

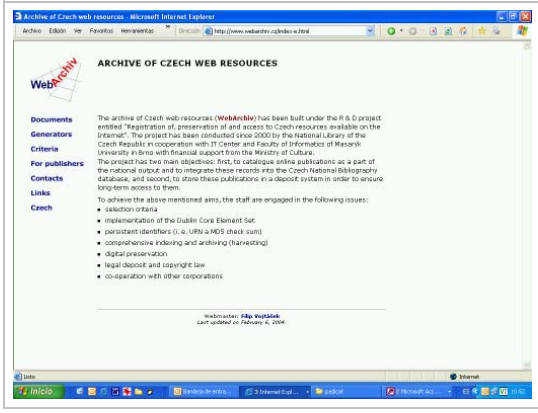


Inici	Lituània, 2002
Lidera	Martynas Mazvydas
Contacte	Remigijus Jodelis (remigijus.jodelis@gmail.com)
Model	Integral
Descripció	El projecte LIBIS Electronic Resources Subsystem consisteix en completar el catàleg LIBIS amb les captures procedents del sistema NEDLIB. El projecte inclou catalogació per Dublin Core.
URL	http://www.inforum.cz/inforum2003/english/prispevek.asp-CisloSekce=9&Kod=44.htm

Paradigma



Inici	Noruega, 2001
Lidera	Nasjonalbiblioteket
Contacte	Carol van Nuys (carol.vannuys@nb.no)
Model	Integral a híbrid
Descripció	Amb captures anuals, a partir de 2003 es va incloure la captura dels dominis internacionals amb contingut noruec, així com 65 diaris digitals. La indexació dels recursos es produeix automàticament pel programari FAST.
URL	http://www.nb.no/paradigma/eng_index.html

Archive of Czech web resources		
	Inici	República Txeca, 2000
	Lidera	Národní knihovna České Republiky
	Contacte	Ludmila Celbova (ludmila.celbova@nkp.cz)
	Model	Integral
	Descripció	En col·laboració amb altres institucions bibliotecàries i de recerca, la biblioteca nacional txeca ha impulsat les captures anuals del domini .cz per mitjà d'una adaptació del programari NEDLIB. És previst incloure publicacions digitals en futures etapes.
	URL	http://www.nb.no/paradigma/eng_index.html

Model selectiu

PANDORA		
	Inici	Austràlia, 1996
	Lidera	National Library of Australia
	Contacte	Margaret Phillips (mphillips@nla.gov.au)
	Model	Selectiu
	Descripció	L'abast del projecte es troba en una selecció de publicacions en línia i webs sobre Austràlia, d'autor australià o sobre un tema australià. La catalogació es exhaustiva, i les possibilitats de cerca són molt avançades. Té un programari propi, PANDAS, que s'ha implementat a altres projectes.
	URL	http://pandora.nla.gov.au/index.html

L'Arxiu Web d'Austràlia, PANDORA, fou creat el 1996 per la Biblioteca Nacional d'Austràlia per garantir l'accés permanent a una selecció de publicacions en línia i seus web de i sobre Austràlia.

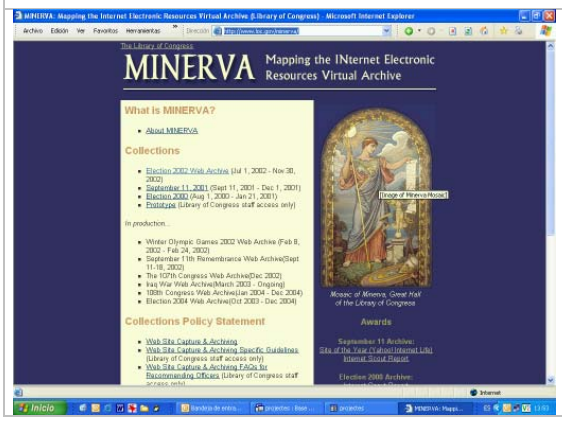
A manca d'una llei que en reguli el dipòsit legal digital (la vigent és de 1968), la política de la Biblioteca i els seus socis de projecte, que formen el comitè *científic* de la política selectiva, és arribar a acords amb les entitats editores dels documents susceptibles de ser capturats. Existeix publicada una guia dels criteris en què es basa la selecció de les seus capturades. Les dades estadístiques (març 2005) presenten 24 milions d'arxius i un creixement mensual de 17 Gb. És consultable en línia.

Els desavantatges[12] del sistema australià estan relacionats amb la seva naturalesa: el criteri de la selecció és forçosament subjectiu, malgrat la transparència de la política que coordini la selecció. El context (els links als quals apunta el recurs), queden deslligats del document, perquè poden no estar inclosos a la selecció. Finalment, el cost de tractament (selecció, captura periòdica, catalogació, etc.) de cada ítem és molt elevat. Per contra, els beneficis es concentren en la qualitat del tractament i presentació del patrimoni. L'accessibilitat en línia, en obert, és possible pels acords subscrits amb els productors (que comporta l'accés als recursos de la Infranet). Les dades de la catalogació són compartibles amb la resta d'equipaments australians (o estrangers). Es procura un creixement temàtic equilibrat de la col·lecció.

E-Collection

Inici	Canadà, 1994
Lidera	Libraries and Archives Canada (LAC)
Contacte	e.publications.e@nlc-bnc.ca
Model	Selectiu
Descripció	A partir de l'EPPP (*Electronic Publications Pilot Project*), de 1994-95, es va crear l'E-Collection, destinat a l'arxiu en línia de publicacions digitals, a text complet. L'actualització del projecte, 2004-05, inclou tesis en línia, webs, etc.
URL	http://epe.lac-bac.gc.ca/

Minerva



Inici	Estats Units d'Amèrica, 2000
Lidera	Library of Congress
Contacte	Martha Anderson
Model	Selectiu temàtic
Descripció	Associat al gegant Internet Archive, el recurs Minerva captura selectivament 35 seus web. En les eleccions presidencials de 2000 (i en altres dates especials: 11S, etc.), s'augmenta la captura selectiva. s previst incloure publicacions digitals en futures etapes.
URL	http://www.loc.gov/minerva/

[12] Phillips, Margaret. "What to collect and how to do it: the National Library of Australia's selective approach", *Archiving web resources: international conference* (Canberra: nov 2004). [en línia]. Canberra: National Library of Australia, 2005. [consulta abril 2005:] http://www.nla.gov.au/webarchiving/program.html

WARP

Inici	Japó, 2002
Lidera	National Diet Library
Contacte	(warp@ndl.go.jp)
Model	Selectiu
Descripció	Les esmenes a la llei de dipòsit legal de 2000 inclouen CDR i altres materials digitals en suports físics. La Biblioteca nacional del Japó recull (juny 2004) amb acords 600 seus web (administració, universitats, empreses, etc.), i 110 diaris electrònics.
URL	http://warp.ndl.go.jp

e-Depot

Inici	Països baixos, 1995
Lidera	Koninklijke Bibliotheek
Contacte	Erik Oltmans (erik.oltmans@kb.nl)
Model	Selectiu
Descripció	A partir dels acords amb els editors (i als Països baixos s'ubiquen un nombre important de multinacionals editores, s'apunta a les revistes publicades a Holanda, on no existeix dipòsit legal de cap mena. 3 milions de números de revistes (març 2005).
URL	http://www.kb.nl/dnp/e-depot/e-depot-en.html

IRIS

Inici	Quebec, 2000
Lidera	Bibliothèque nationale du Québec
Contacte	Maureen Clapperton (pubelectro@bnquebec.ca)
Model	Selectiu
Descripció	3.469 monografies i 1.100 títols de revistes digitals (octubre 2004) formen el cos de l'arxiu, que està integrat al catàleg IRIS. Captura amb acords amb la pròpia Administració en una primera fase.
URL	http://catalogue.bnquebec.ca:4400/cap_fr.html

UK Web Archive

Inici	Regne Unit, 2004
Lidera	British Library
Contacte	Mark Middleton (mark.middleton@bl.uk)
Model	Selectiu
Descripció	Seguint el model australià, amb el mateix programari PANDAS i una aparença molt similar , el maig de 2005 presentava 1030 seus web ab les que s'ha arribat a acords. En la selecció, i en tot el projecte, participen la resta de biblioteques nacionals del Regne Unit.
URL	http://www.webarchive.org.uk/

Model híbrid

Com s'ha apuntat, bona part dels repositoris nacionals plantejats com integrals han anat adoptant mesures per incloure recursos molt significatius (publicacions periòdiques, etc.) als seus fons.

En tot cas, tres són els projectes pioners en apostar per una política clara de conjugació de les captures exhaustives, els acords amb institucions i organitzacions, i el detall per a activitats concretes: França, Dinamarca, i Nova Zelanda.

Es reprodueix el gràfic que representa l'abast del model danès:

Model híbrid:
el cas de Dinamarca

Captura exhaustiva de la web danesa (*.dk)
+
Acords amb entitats editores (administració, universitats, etc.)
+
Captura integral d'events específics (eleccions, esports, etc.)

Netarkivet		
	Inici	Dinamarca, 1998
	Lidera	Det Kongelige Bibliotek
	Contacte	Bjarne Andersen (bja@netarkivet.dk)
	Model	Híbrid
	Descripció	El més difós dels models híbrids, basat en la captura exhaustiva, els acords per selecció, i les activitats especials relacionades amb la realitat danesa. Des de 1997 la llei de dipòsit legal inclou "totes" les publicacions de Dinamarca. La biblioteca nacional danesa inclou el lliurament del dipòsit legal per mitjà d'un formulari web.
	URL	http://netarchive.dk/index-en.php

A partir d'un model inicial (domini .dk), el 2004 s'adopta el sistema híbrid, adreçat com s'ha esmentat al triple objectiu (integral+selectiu+especials). Hi participen la Kongelige Bibliotek (Biblioteca Nacional de Dinamarca) i la State and University Library (Uhus). Puntualment (selecció seus web de literatura danesa) s'hi incorporen entitats temàticament vinculades.

La mida del Netarkivet ronda (2004) els 500 Tbyte, amb un exponent de creixement anual de 30 Tb. No és accessible en línia.

El projecte danès té punts forts evidents, que són infraestructurals:

- Dinamarca és un país petit (5,4 M. habitats), amb un domini propi (dk) i llengua pròpia (danès), la selecció (i les gestions conseqüents) és relativament senzilla per limitada.
- Per la mateixa raó, també són senzilles les captures exhaustives. Per al desenvolupament del programari, el Netarkivet ha col·laborat al Nordic Web Index (NWI), amb la resta de biblioteques escandinaves.

- La legislació que afecta el dipòsit legal ha anat modificant-se (darrera revisió, a començaments de juliol de 2005[13]), per ampliar-se a les publicacions digitals en línia. Un formulari en línia facilita la tasca dels productors i editors web.

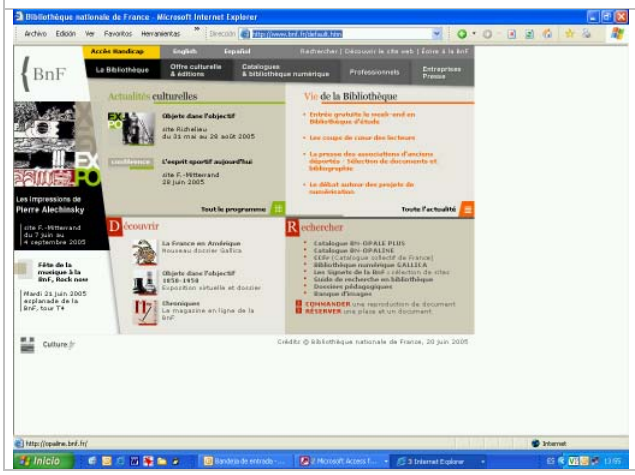
- Dos importants socis estan involucrats en la coordinació del projecte: la biblioteca nacional i un centre de recerca d'una important universitat. S'elaboren acords estratègics en base a determinades temàtiques.

Malgrat aquests fets, el desenvolupament del projecte ha estat irregular o ho aparenta per la bibliografia existent.

El 24 de juny[14] d'enguany s'anuncia al gran públic la posada en marxa, lligant-ho a la nova modificació de la llei de dipòsit legal, del projecte danès. Però amb anterioritat s'havia també impulsat dues fases (2001, 2003), amb idèntica intensitat, i el fet que no sigui accessible en línia és un punt feble que una entitat cultural nacional ha d'impulsar, ni que sigui parcialment.

El juliol de 2005 s'inicia la primera fase de captura global, amb el programari Heritrix.

Archiving the French web



Inici	França, 2000
Lidera	Bibliotheque Nationale de France (BNF)
Contacte	Catherine Lupovici (catherine.lupovici@bnf.fr)
Model	Híbrid
Descripció	L'abast del projecte inclou captures automàtiques a gran escala, captures sistemàtiques i contínues d'una selecció de seus web (el 10% del total), dipòsit de la Infranet, i captures temàtiques de seus web molt efímeres (eleccions franceses de 2002: 1900 seus web).
URL	ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-229.pdf

[13] "New legal deposit law", en *Netarchive.dk* [consulta juliol 2005:] http://netarchive.dk/newsite/news/index-en.php
[14] La data és simbòlica: la festa de sant Joan (el solstici d'estiu) és especialment festiva als països escandinaus.

New Zealand's digital Heritage		
	Inici	Nova Zelanda, 1999
	Lidera	National Library of New Zealand
	Contacte	Shelley Cartwright (shelley.cartwright@natlib.govt.nz)
	Model	Híbrid
	Descripció	La llei de dipòsit legal novazelandesa, de 2003, obria el panorama als recursos en línia, incloent les publicacions en obert, i també la Infranet. Pressupost global patrimoni digital (2004): 14 M€. Empra el programari PANDAS, però l'accés no és en obert.
	URL	http://www.natlib.govt.nz/bin/media/pr?item=1085885702

2.4. Organitzacions i projectes suprainstitucionals

International Internet Preservatium Consortium

L'organització que aplega la major part d'aquestes iniciatives és l'International Internet Preservatium Consortium[15] (IIPC), que té la missió d'adquirir, preservar i fer accessible el coneixement i la informació sobre Internet per a les futures generacions de tot el món, promovent l'intercanvi global i les relacions internacionals.

Creat formalment el juliol de 2003 pels 12 membres que actualment formen el consorci: Bibliothèque Nationale de France (coordinador), Biblioteca Nazionale Centrale di Firenze (Itàlia), Det Kongelige Bibliotek (Dinamarca), Helsingin yliopiston kirjasto-Suomen Kansalliskirjasto (Finlàndia), Internet Archive (EUA), Kungliga biblioteket Sveriges nationalbibliotek (Suècia), Landsbokasafn Islands- Haskolabokasafn (Islàndia), Library and Archives Canada, Nasjonalbiblioteket (Noruega), National Library of Australia, The British Library, i The Library of Congress (EUA).

És previst incorporar nous membres (2006), i en aquest sentit s'ha contactat amb la coordinació (Catherine Lupovici, BNF) del Consorci.

Els objectius del Consorci són:

- Permetre la recol·lecció d'una part rica del contingut d'Internet d'arreu del món, per a ser preservada de forma que pugui ser arxivada, preservada i assegurat l'accés en el temps.

[15] *International Internet Preservatium Consortium* [en línia]. [S.l.]: IIPC, 2005 [consulta juny 2005:] http://netpreserve.org/

- Fomentar el desenvolupament i ús d'eines comunes, tècniques i estàndards que permetin la creació d'arxius internacionals.

- Animar i donar suport a les biblioteques nacionals d'arreu per arxivar i preservar Internet.

Existeixen diversos grups de treball (eines d'accés, gestió de continguts, etc.) creats a l'empara del Consorci, i la intenció de publicar informes i facilitar l'accés a programari, qüestions aquestes que no han estat públicament completades.

El Consorci, doncs, no captura webs, però sí agrupa una sèrie d'institucions que sí ho fan, i n'és objectiu la promoció d'aquestes activitats.

Internet Archive

L'Internet Archive[16] és una organització sense ànim de lucre fundada el 1996 per construir una "biblioteca d'Internet", i oferir accés permanent per a investigadors, historiadors, i escolars a les col·leccions històriques que existien en format digital. Situada a l'antiga presó de San Francisco (EUA), l'arxiu ha rebut donacions d'IBM, d'Alexa (filial d'Amazon) i altres recursos similars, que han facilitat el seu creixement.

Diversos socis donen suport al recurs, com la Library of Congress, els US National Archives, els UK National Archives, entre d'altres.

A l'actualitat Internet Archive es considera l'arxiu web més gran del món[17], i inclou text, àudio, imatge en moviment, i programari, així com pàgines web arxivades de tot el món, incloent-hi un bon número de recursos catalans[18] que PADICAT podrà aprofitar per al fons retrospectiu. El recurs conté en accés obert en línia 35 milions de seus web, des de 1996 fins a l'actualitat, i cada dos mesos es realitza una captura massiva que afecta 4.000 milions de pàgines web, seguint el model exhaustiu que Suècia i altres països representen.

El programa Heritrix és el gestor (programari lliure) que utilitza l'Internet Archive, i el sistema d'emmagatzematge es realitza en múltiples còpies, separades geogràficament.

---

[16] *Internet Archive* [en línia]. San Francisco: Internet Archive, 2005. [consulta juny 2005:] http://www.archive.org

[17] Kimpton, Michele. "Saving the web for future generations", *Archiving web resources: international conference* (Canberra: nov 2004). [en línia]. Canberra: National Library of Australia, 2005. [consulta juny 2005:] http://www.nla.gov.au/webarchiving/MicheleKimpton.ppt

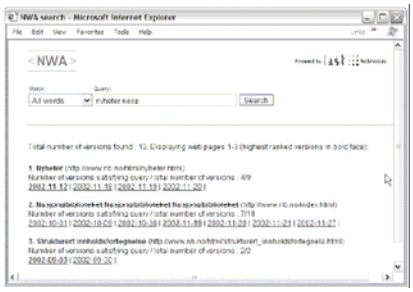
[18] Alguns exemples són: www.gencat.net (207 captures des de maig de 2002), www.avui.es (55 captures des de gener de 1998), www.ub.edu i www.ub.es (223 captures des de febrer de 1997), etc.

Recentment, l'antic coordinador de l'International Internet Preservatium Consortium, Julien Masanès, s'ha incorporat a dirigir l'European Internet Archive, la branca europea de l'Internet Archive.

Nordic Web Archive – NEDLIB

El Nordic Web Archive[19] (NWA) és de fet un fòrum de les biblioteques nacionals escandinaves (Dinamarca, Finlàndia, Islàndia, Noruega i Suècia) per a la coordinació i intercanvi d'experiències en els camps de la captura i arxiu de documents web[20].

Des de novembre de 2000 s'ha desenvolupat el set d'eines NWA: un paquet de programari per accedir als document web arxivats, creat emprant PHP, Perl i Java, amb estàndards oberts com el protocol http i XML per la comunicació entre les diferents parts del sistema. L'ús del paquet de programari (cerca i navegació per l'arxiu web), es realitza per mitjà d'un cercador web estàndard, i cap *plugin* específic és necessari.



L'activitat va ser fundada per Nordunet2 (Programa de recerca dels escandinaus), Nordinfo (Consell escandinau per a la Informació Científica que inclou les biblioteques de recerca), i les biblioteques nacionals escandinaves, i el paquet de programari és descarregable a Sourceforge.

2.5. Conclusions sobre l'estat de la qüestió arreu del món

- L'interès per la preservació digital està ja generalitzat als països desenvolupats, encara que amb una terminologia encara diversa (arxiu web, repositori nacional, patrimoni digital). Possiblement a data d'avui existeixin més dels vint projectes mencionats, encara que siguin en fase de disseny.

- El futur és híbrid: la diferenciació en models (integral vs. selectiu) és només una primera fase. Dinamarca, Austràlia, i en darrer terme Suècia, són els models a seguir.

- Els projectes de repositori són econòmicament costosos, i passen forçosament per la implicació del nombre més elevat d'agents possibles, que dotin de continuïtat als programes un cop endegats. En aquest sentit, els fracassos que regularment afecten els projectes estudiats ho són per manca de finançament.

[19] *Nordic Web Archive* [en línia]. Oslo: National Library or Norway, 2004. [consulta juny 2005:] http://nwa.nb.no/
[20] Hallgrímsson, Porsteinn; Bang, Sverre. "Nordic Web Archive", *International Web Archiving Workshop* (Trondheim : 2003). [en línia]. [S.l.: s.n.], 2003. [consulta juny 2005]: http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf

- Existeix un corrent global de cooperació (compartir experiències, el relat dels èxits i fracassos, programari en codi obert) entre els projectes. L'exemple més evident és la generalització del programa Heritrix.

- Els acords amb els productors i editors web són garantia d'èxit. No sempre una llei moderna de dipòsit legal acompanya als repositoris nacionals que existeixen, i l'accés a la Infranet (i la Internet invisible) ha de contemplar-se.

3. Context a Catalunya: recursos existents, agents implicats, aspectes legals

3.1. Introducció

L'anàlisi del context a Catalunya és bàsic per abordar amb garanties un model de dipòsit nacional, l'híbrid[21], que contempla els acords de la BC amb les entitats que per la seva pròpia activitat produeixen continguts digitals.

De la capacitat de lideratge de la BC dependrà la participació dels agents implicats per garantir la màxima cooperació.

Per citar exemples destacats dels fruits de la cooperació: la creació de catàlegs bibliogràfics compartits (BEG, CCUC, etc.), dels que ja la BC forma part; o el fet el Consorci de Biblioteques Universitàries de Catalunya (CBUC) ha impulsat la Biblioteca Digital de Catalunya (accés obert bases de dades) i el TDX (Tesis doctorals en xarxa), que està en la línia encetada a altres països, com el Regne Unit o Austràlia. El darrer projecte relacionat amb aquesta premissa, en el qual participen la BC, el CBUC i el CESCA, és el RACO (Revistes Catalanes amb Accés Obert), dipòsit de revistes científiques, culturals i erudites catalanes. La conclusió és que sintonitzar els esforços de diversos equipaments punters de la societat fan viable un projecte ambiciós de preservació del patrimoni digital.

No es coneixen accions similars a la resta de l'Estat.

Podem considerar *agents* de la producció digital[22] els propis autors, editorials, universitats, intermediaris diversos, biblioteques, usuaris, i la resta d'actors que, aprofitant les avantatges de les TIC, creen i mantenen seus web de diversa índole i format (administració –*eGovernment*--, televisió digital, indústria discogràfica, etc.). Són els agents susceptibles de ser implicats en el projecte PADICAT, i la dificultat de preveure el tracte personalitzat que hauran de rebre no ha d'obstaculitzar la previsió d'uns models a partir de la naturalesa d'aquestes institucions.

De la capacitat de lideratge de la BC dependrà el nivell de participació dels agents involucrats per assegurar-ne la màxima cooperació. La fórmula cessió x preservació *(cessió* de l'agent productor = *tractament i preservació garantit* del centre nacional) s'utilitza actualment a diversos projectes, amb resultats exponencialment positius.

[21] En l'informe *Estat de la qüestió arreu del món* (versió 1.1 setembre 2005) del projecte es presenta les diferents estratègies que s'han adoptat per dur a terme projectes d'arxiu web. A partir dels dos models inicials (selectiu i integral), la tendència és l'adopció generalitzada del model híbrid (Dinamarca, França, i Nova Zelanda), que consisteix a realitzar captures sistemàtiques de la web nacional, arribar a acords amb els productors de la producció digital del territori, i fer captures monogràfiques al voltant d'esdeveniments concrets. A manca d'una anàlisi detallada del programari i la capacitat de finançament de la BC, el model híbrid és per tant la referència del PADICAT.
[22] Keefer, Alice. *Preservació dels recursos d'informació digital* [electrònic]. Barcelona: UOC, 2003.

Pel que afecta als aspectes legals relacionats amb la captura i tractament d'informació digital, l'obsolescència del text vigent, de 1971[23], fa que calgui començar de forma immediata la recopilació del patrimoni digital, prescindint de l'*alegalitat* de l'acció. Si finalment es regula i la captura resulta prohibida (cas extrem d'imprevisió dels òrgans legisladors), la institució que hagi avançat terreny podrà decidir-ne llavors el destí dels seus dipòsits digitals. Si com cal esperar se'n protegeix i s'estimula la conservació i l'accés permanent al patrimoni digital, la institució que s'hagi anticipat, com és el cas de la BC, gaudirà d'un avantatge competitiu que revertirà en inversions i suport polític. Així, els aspectes legals relacionats amb la pràctica objectiu del PADICAT han de ser mesurats tenint en compte la realitat catalana i espanyola, amb un text legal desfasat i sense perspectiva de canvi a curt termini, així com també l'acció realitzada en altres països amb projectes de dipòsit digital sistematitzat, que han hagut d'adaptar-se a la llei existent, o fins i tot han pogut contemplar l'adequació del dipòsit legal a la nova realitat.

De fet, la funció de la Biblioteca de Catalunya *de recollir les obres editades o produïdes a Catalunya*[24] forma part de la legislació catalana, que en l'abstracció del redactat no distingeix entre obres impreses o en suport digital. Paral·lelament, una tasca pedagògica de pressió als legisladors hauria de perseguir l'actualització de la legislació del dipòsit legal cercant uns màxims (model suec, competència exclusiva de la biblioteca nacional) a partir d'uns mínims (preservació per llei del patrimoni digital, diversos òrgans implicats). Finalment, amb suport legal o sense, la voluntat de les entitats productores (universitats i centres de recerca, empreses i gremis, administració, col·legis professionals, associacions, partits polítics i sindicats, particulars, etc.) ha de veure el PADICAT com l'oportunitat de ser els protagonistes de la futura recerca, així com de formar part del Patrimoni Digital de Catalunya, entès com un sistema útil per a la societat i les institucions que la formen.

Ni a Catalunya ni a Espanya es té constància d'accions similars al PADICAT, però és probable que en els propers mesos apareguin programes adreçats a compilar la web catalana o espanyola en base a la producció d'una universitat, sobre un determinada temàtica, etc.

3.2. Recursos existents

Ni a Catalunya ni a Espanya es té constància d'accions similars[25] a les que són objectiu del projecte PADICAT.

De fet, si es parteix de la definició estricta del dipòsit nacional difícilment es podrà trobar paral·lelisme amb algun projecte espanyol. Evidentment però, la inquietud existeix[26] i és probable que en els propers

[23] "Orden de 30 de octubre de 1971, por la que se aprueba el Reglamento del Instituto Bibliográfico Hispánico", *Boletín Oficial del Estado*, (18 nov 1971). En el capítol II es regula el Dipòsit Legal. De fet, existeix un decret de 1958 que és el precedent del text de 1971.
[24] Article 9.1 de la "Llei 4/1993, de 18 de març, del sistema bibliotecari de Catalunya", en *Diari Oficial de la Generalitat de Catalunya* (29 mar 1993).
[25] Agenjo, Xavier; Hernández, Francisca. "La recolección de metadatos (metadata harvesting) y su aplicación en España", en *Infogestión: IX Jornadas Españolas de Documentación Fesabid 2005*. Madrid: Sedic, 2005.

mesos apareguin programes adreçats a compilar la web catalana o espanyola en base a determinats temes (la producció pròpia d'una universitat --la UCM és actualment l'únic cas--, en base a una temàtica –portal Temaria de la UB--, sobre un determinat escriptor o zona geogràfica, etc.). De fet, la *Subdirección General de Coordinación Bibliotecaria* té un projecte destinar a crear un dipòsit de diverses col·leccions digitals, essent el referent el model MINERVA, creat als Estats Units amb propòsit temàtic.

Consultades les universitats públiques catalanes[27] per mitjà de les seves biblioteques, és factible afirmar que hi ha previsió de crear dipòsits digitals institucionals (sistema Dspace), però aquest procés és encara en fase molt incipient. S'avançarà la UPC, que el proper octubre preveu fer públics els repositoris de revistes i d'e-prints[28].

Com s'ha comentat a la introducció, i pel que fa estrictament a la cooperació entre institucions, els exemples importants a Catalunya (ho són també a la resta de l'Estat) són més nombrosos: els catàlegs compartits (BEG, CCUC, SLP, XPB, etc. en procés d'unificació), les Tesis Doctorals en Xarxa[29] (TDX) i la Biblioteca Digital de Catalunya[30] (BDG), impulsades pel Consorci de Biblioteques Universitàries de Catalunya (CBUC), i en darrera instància el Revistes Catalanes amb Accés Obert[31] (RACO), del mateix consorci, el CESCA i la Biblioteca de Catalunya, amb el suport del DURSI.

S'especificarà en la descripció detallada dels recursos, però avancem que aquests projectes no són pròpiament dipòsits digitals dels documents *nascuts digitals* (si exceptuem les publicacions del RACO que compleixen aquesta característica). Sí són mostra de la tasca ingent, però possible, que és necessària en la implicació dels agents productors per tal de garantir la creació i manteniment de recursos digitals a l'abast de la societat.

---

[26] El proper mes de març se celebren a Madrid unes jornades dedicades a la preservació digital internacional, organitzades pel Grup de Treball de patrimoni digital de la Subdirección General de Cooperación Bibliotecaria del Ministerio de Cultura.
[27] UAB Núria Balagué; UB Montse Playá; UdG Antònia Boix; UDL Loli Manciñeiras; UOC Dora Pérez (i Mireia Riera IN3, programa conjunt Universitas 21) ; UPC Marta López-Vivancos (la responsable és Anna Rovira); UPF Roger Esparó; URV Mariantònia Aloguín.
[28] Revistes (https://eprints.upc.es:8443/revistes) i E-prints (https://eprints.upc.es:8443/dspace)
[29] El Servidor de Tesis Doctorals en Xarxa (TDX) conté, en format digital, tesis doctorals llegides a les universitats de Catalunya i d'altres comunitats autònomes. Permet la consulta remota a través de la xarxa Internet del text complet de les tesis, així com fer cerques per autor, títol, matèria de la tesi, universitat on s'ha llegit, etc. El servei està gestionat pel Consorci de Biblioteques Universitàries de Catalunya (CBUC) i el Centre de Supercomputació de Catalunya (CESCA), i patrocinat pel Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya.
[30] La Biblioteca Digital de Catalunya (BDC) és un recull Biblioteca Digital de Catalunya (BDC) és un recull d'informació electrònica (revistes, bases de dades i llibres electrònics) subscrita conjuntament per tots els membres del CBUC i consultable des de qualsevol punt autoritzat de la xarxa, mitjançant control d'adreça IP. Part d'aquesta informació (les bases de dades catalanes) és oberta a tothom.
[31] RACO (Revistes Catalanes amb Accés Obert) és un portal en fase de desenvolupament des del qual es poden consultar en accés obert els articles a text complet de més de 40 revistes científiques, culturals i erudites catalanes. La finalitat de RACO és augmentar la visibilitat i consulta de les revistes que inclou i difondre la producció científica i acadèmica que es

A Catalunya:

Recurs	Catàleg Col·lectiu de les Universitats de Catalunya (CCUC)[32]
Impulsa	Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes --UAB, UB, UDG, UDL, UOC, UPC, UPF, URV--, la Biblioteca de Catalunya, i el DURSI)
Objectius	El CBUC té per missió la de millorar els serveis bibliotecaris a través de la cooperació.
Conté	2.500.000 registres bibliogràfics (representen uns 5 milions de documents físics de 145 biblioteques), i accés als recursos electrònics de la Biblioteca Digital de Catalunya.
Característiques	Catàleg a partir del que es formen els diferents catàlegs locals per mitjà d'un mecanisme de lligams, regit per un sistema de control de qualitat i les directrius que marquen les institucions que formen el Consorci.

Recurs	Tesis Doctorals en Xarxa (TDX)
Impulsa	Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes --UAB, UB, UDG, UDL, UOC, UPC, UPF, URV--, la Biblioteca de Catalunya, i el DURSI) i el Centre de Supercomputació de Catalunya (CESCA), amb el patrocini del DURSI.
Objectius	Difondre arreu del món, per Internet, els resultats de la recerca universitària, oferir als autors de les tesis una eina que incrementa l'accés i la visibilitat del seu treball, millorar el control bibliogràfic de les tesis, impulsar l'edició electrònica i les biblioteques digitals, incentivar la creació i l'ús de la producció científica pròpia.
Conté	2.700 tesis doctorals llegides a les universitats de Catalunya i d'altres comunitats autònomes (14 en total: 9 universitats consorciades i 5 externes).
Característiques	Permet la cerca i consulta a través d'Internet al text complet de les tesis.

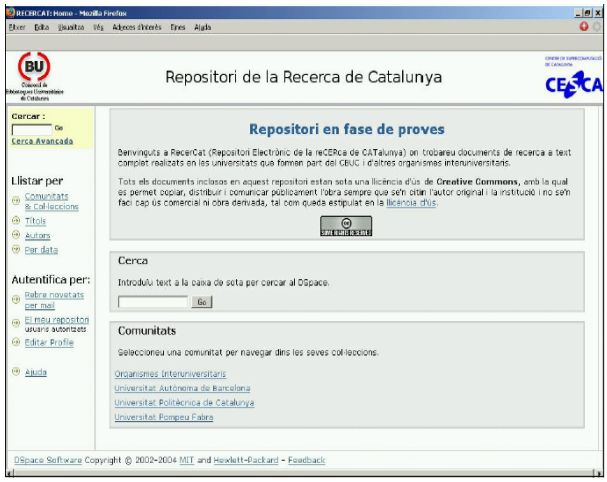
Recurs	Biblioteca Digital de Catalunya (BDC)
Impulsa	Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes --UAB, UB, UDG, UDL, UOC, UPC, UPF, URV--, la Biblioteca de Catalunya, i el DURSI), amb el suport del DURSI.
Objectius	L'objectiu és oferir un conjunt nuclear d'informació electrònica interdisciplinar per a la totalitat de la comunitat universitària i investigadora de Catalunya, independentment d'on aquestes persones exerceixin aquesta activitat.
Conté	Recull seleccionat d'informació en format electònic (a març de 2005: 6.800 revistes, 58 bases de dades i 4.100 llibres digitals) subscrita conjuntament pels membres del Consorci.
Característiques	Consultable majoritàriament des dels punts autoritzats (per mitjà del control dels IP dels membres del Consorci) de la xarxa.

Recurs	Revistes Catalanes amb Accés Obert (RACO)
Impulsa	Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes --UAB, UB, UDG, UDL, UOC, UPC, UPF, URV--, la Biblioteca de Catalunya, i el DURSI), la Biblioteca de Catalunya, i el CESCA, amb el suport del DURSI.
Objectius	La finalitat és augmentar la visibilitat i consulta de les revistes que inclou el RACO i difondre la producció científica i acadèmica que es publica a revistes catalanes, en són objectius impulsar l'edició electrònica de revistes catalanes, crear una interfície que permeti la consulta conjunta de totes les revistes, i facilitar els instruments per la seva preservació.
Conté	Consulta i accés obert als articles de més de 40 revistes científiques, culturals i erudites catalanes.
Característiques	Consultable en obert, procura l'accés a la versió digital (o digitalitzada) de les revistes que han subscrit l'acord. Té un servei d'alerta.

publica a revistes catalanes, en són objectius impulsar l'edició electrònica de revistes catalanes, crear una interfície que permeti la consulta conjunta de totes les revistes, i facilitar els instruments per la seva preservació.

[32] El CCUC no és un dipòsit digital, però sí l'exemple més evident de la força de la cooperació entre agents.

Recurs	Repositori de la Recerca de Catalunya (RECERCAT)[33]
Impulsa	Consorci de Biblioteques Universitàries de Catalunya (CBUC: format per les universitats públiques catalanes --UAB, UB, UDG, UDL, UOC, UPC, UPF, URV--, la Biblioteca de Catalunya, i el DURSI).
Objectius	La finalitat és crear un repositori institucional de les universitats de són membres, per tal de facilitar l'edició (entrada autoalimentada de dades), augmentar la visibilitat (metadades, llicències Creative Commons), afegir valor als documents (citacions normalitzades, adreces permanents, preservació, etc.) de la literatura de recerca no publicada i generada a les universitats de Catalunya.
Conté	En fase de proves.
Característiques	Consultable en obert, procurarà l'accés als articles dels centres que han subscrit l'acord.



A Espanya:

Recurs	Biblioteca Virtual de Andalucía (BVA)
Impulsa	Junta de Andalucía
Objectius	Efectuar la digitalització d'originals (llibres, mapes, manuscrits, etc.) amb difícil accés, localitzats en diferents institucions culturals, dins i fora d'Andalusia; difusió del fons bibliogràfic; accés al patrimoni bibliogràfic andalús *nascut digital*.
Conté	Conjunt de col·leccions de documents digitalitzats del patrimoni bibliogràfic andalús.
Característiques	Accessible en obert per mitjà d'Internet.

Recurs	E-Prints de la Universidad Complutense de Madrid (E-PrintsUCM)
Impulsa	Universidad Complutense de Madrid
Objectius	Compilar i oferir en obert la producció digital dels docents i investigadors de la UCM.
Conté	3.285 documents (word, pdf, html)
Característiques	Accessible en obert per mitjà d'Internet.

Recurs	Tecnociencia e-revistas (e-revist@s)[34]
Impulsa	Fundación Española de Ciencia y Tecnología (FECYT)
Objectius	Creació d'una plataforma digital on es compilin, seleccionin i hostatgin revistes electròniques espanyoles o llatinoamericanes existents amb criteris de selecció qualitatius.
Conté	3.418 articles de les 56 revistes amb les que s'ha arribat a acords (un 70% del total) amb una temàtica multidisciplinar: ciència i tècnica.
Característiques	Cerca i text complet accessible en obert per mitjà d'Internet.

[33] Anglada, Lluís; Miquel Huguet. "Selecció del programa Dspace per a la gestió d'un repositori de documents de recerca per les Universitats de Catalunya" [en línia], en Internet Global Conference (7è: Barcelona: 2005). [consulta agost 2005:] http://www.cesca.es/promocio/conferencies/2005/0506IGC.PDF

[34] Fernández, Elena; Luis Rodríguez; Juan Francisco Heras. "La plataforma e-revist@s del portal Tecnociencia: una experiencia basada en open access", en *El profesional de la información*, v. 14, n. 4 (jul-ago 2005), p. 290-296.

Recurs	Biblioteca Virtual Miguel de Cervantes (Cervantes Virtual)
Impulsa	Fundación Biblioteca Virtual Miguel de Cervantes[35] (creada per la Universitat d'Alacant, el Grup bancari Santander i la Fundació Marcelino Botín, s'hi ha afegit nous membres.)
Objectius	Va néixer 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.
Conté	Col·lecció de materials digitals (*nascuts* i digitalitzats), però Cervantesvirtual també és un centre d'investigació i un vehicle d'extensió de la cultura hispànica.
Característiques	Cerca i text complet accessible en obert per mitjà d'Internet.

Com s'ha esmentat a peu de pàgina, s'ha previst realitzar diversos contactes amb les institucions impulsores dels projectes per conèixer de prop les polítiques en la selecció i acords dels dipòsits de documents *nascuts digitals*, així com en la infraestructura que s'utilitza en la provisió d'accés, en ambdós casos.

La conclusió és que sintonitzar els esforços de diversos equipaments punters de la societat fan viable un projecte ambiciós de preservació del patrimoni digital. La selecció i l'arribada estratègica a acords amb els productors poden crear un efecte cascada que sigui beneficiós per al PADICAT. Es tracta, en primer lloc, d'analitzar quins són els agents implicats.

3.3. Agents implicats

Podem considerar *agents* de la producció digital[36] els propis autors, editorials, universitats, intermediaris diversos, biblioteques, usuaris, i la resta d'actors que, aprofitant les avantatges de les TIC, creen i mantenen seus web de diversa índole i format (administració –*eGovernment*--, televisió digital, indústria discogràfica, etc.). Són els agents susceptibles de ser implicats en el projecte PADICAT, i la dificultat de preveure el tracte personalitzat que hauran de rebre no ha d'obstaculitzar la previsió d'uns models a partir de la naturalesa d'aquestes institucions.

No es pot subestimar la inversió en hores de personal que suposa la necessària gestió dels acords amb els agents implicats en la producció digital catalana. En el projecte danès representa el 44% del pressupost total.

Com s'ha apuntat a la introducció, de la capacitat de lideratge de la BC dependrà el nivell de participació dels agents involucrats[37] per assegurar-ne la màxima cooperació. La fórmula cessió x preservació *(cessió* de l'agent productor = *tractament i preservació garantit* del centre nacional) s'utilitza actualment a diversos projectes europeus, amb resultats exponencialment positius.

[35] S'ha contactat amb Julia Bernal i Laura Sánchez, subdirectora i gerent respectivament de la Cervantes Virtual, per realitzar una visita a les instal·lacions i conèixer amb més detall la infraestructura i els processos.
[36] Keefer, Alice. *Preservació dels recursos d'informació digital* [electrònic]. Barcelona: UOC, 2003.

Recopilar un llistat d'agents implicats comporta en certa manera seleccionar quins són els objectius conceptuals del projecte, quines són les prioritats, i tanmateix aprofundir en definir quins són els agents d'interès. Els que més produeixen? Els que produeixen informació nascuda digital de més qualitat i rellevància?

De fet, les desavantatges típiques dels dipòsits digitals basats en el model selectiu[38] inclouen el judici subjectiu sobre el que els investigadors requeriran en el futur; la pèrdua inevitable de recursos importants; la labor intensiva (i el cost elevat) d'aquesta selecció; la separació dels recursos del seu context, etc.

D'altra banda, és temptador seleccionar els agents implicats en base al format de la seva producció associat a l'agent productor, com s'ha realitzat a Pandora[39]:

Una primera selecció focalitzada en categories, basada en el cas australià, contemplaria:

- Publicacions de la Generalitat de Catalunya
- Publicacions de les Universitats i centres d'educació superior i recerca
- Actes de conferències
- Revistes digitals
- Ítems creats en pels agents creadors d'índexs i resums
- Webs relacionades amb les temàtiques de la selecció d'esdeveniments concrets (esports, eleccions, etc.)

Seguint el cas australià, no es contemplaria:

- Publicacions diàries en línia
- Seus web de notícies
- Llistes de distribució, Xats, butlletins i grups de notícies
- CAMs
- Blogs (excepte el que donguin suport a les publicacions acadèmiques)
- Jocs en línia

Però el fet d'arribar a acords amb agents productors en la publicació web catalana hauria de garantir el dipòsit digital dels materials que es produeixin en aquestes entitats, indiferentment del format de la producció. De fet, arribar a acords amb aquests agents servirà per arribar a la web invisible (que queda protegida per les contrasenyes, pàgines dinàmiques, etc.), i tanmateix per possibilitar l'accés públic a la

[37] N'és exemple l'acord de cooperació subscrit el 2002 entre la Biblioteca Alemanya i l'Associació del Comerç del Llibre Alemany pel qual es facilita el dipòsit voluntari de publicacions als associats, en sintonia amb el moviment *Arxiu Obert* de la comunitat universitària mundial.

[38] Com el model australià, des d'on s'han extret aquests punts negatius.

[39] Phillips, Margaret E. *Colleting Australian online publications*. [en línia]. Canberra: National Library of Australia, 2003. [consulta juliol 2005:] http://pandora.nla.gov.au/bsc49.doc

informació, amb les limitacions que cada organització vulgui formalitzar en els acords amb la Biblioteca de Catalunya.

Per a cadascun dels agents, l'objectiu és triple:

- Aconseguir la compilació exhaustiva d'aquells agents especialment dinàmics, que modifiquen la seva web amb una regularitat que la captura massiva (anual, semestral) no podria captar.

- Aconseguir publicar en obert la producció (web, publicacions, etc.) del màxim nombre d'agents, amb les limitacions temporals pertinents d'accés a la part oculta de les webs.

- Aconseguir un creixement equilibrat del fons Padicat, amb la incorporació de les webs d'institucions que no ofereixen un volum important de continguts, però sí una mostra transversal del que és la web catalana, actualment.

Proposem a continuació un llistat inicial orientatiu d'agents a desenvolupar en tres fases (A, prioritat alta; B, mitja; C, baixa) en correspondència als recursos que es destinin per al la gestió dels acords, i tanmateix en base a la temàtica que representen i el tipus d'informació que publiquen, per garantir un creixement equilibrat en contingut i forma. En el procés normalitzat de selecció i captura de recursos és imprescindible la participació d'un comitè científic procedent de les institucions implicables.

Agent	*Fase*
Generalitat de Catalunya	A
President de la Generalitat	A
Generalitat de Catalunya. Departament de la Presidència (inclou el Consell Català de l'Esport; l'Entitat Autònoma del Diari Oficial i de Publicacions de la Generalitat de Catalunya; l'Institut Català de la Dona; i el Patronat Català Pro-Europa)	A
Generalitat de Catalunya. Departament d'Agricultura, Ramaderia i Pesca	B
Generalitat de Catalunya. Departament de Benestar i Família	B
Generalitat de Catalunya. Departament de Comerç, Turisme i Consum (inclou el Consorci de Promoció Comercial de Catalunya, COPCA)	B
Generalitat de Catalunya. Departament de Cultura (inclou els equipaments culturals consorciats: Biblioteca de Catalunya, etc.)	A
Generalitat de Catalunya. Departament d'Economia i Finances (inclou l'Institut Català de Finances; l'Institut d'Estadística de Catalunya)	A
Generalitat de Catalunya. Departament d'Educació (inclou la Xarxa Telemàtica Educativa de Catalunya, XTEC)	A
Generalitat de Catalunya. Departament de Governació i Administracions públiques	B
Generalitat de Catalunya. Departament d'Interior (inclou el Servei Català de Trànsit)	B
Generalitat de Catalunya. Departament de Justícia	B
Generalitat de Catalunya. Departament de Medi ambient i Obres públiques (inclou el Servei Meteorològic de Catalunya)	B
Generalitat de Catalunya. Departament de Política territorial i Obres públiques (inclou l'Institut Cartogràfic de Catalunya)	B
Generalitat de Catalunya. Departament de Relacions institucionals i participació	B
Generalitat de Catalunya. Departament de Salut (inclou el Servei Català de la Salut)	B
Generalitat de Catalunya. Departament de Treball i Indústria	B

(inclou el Centre d'Innovació i Desenvolupament Empresarials, CIDEM)	
Generalitat de Catalunya. Departament d'Universitats, Recerca i Societat de la Informació	A
Agència Catalana de Protecció de Dades	B
Corporació Catalana de Ràdio i Televisió	B
Parlament de Catalunya	B
Consell de l'Audiovisual de Catalunya	B
Sindicatura de Comptes	B
Síndic de Greuges	B
Administració de l'Estat dins de Catalunya	C
Poder judicial (inclou audiències provincials, deganats, fiscalies, juntes electorals, jutjats i tribunals)	C
Administració local	B
Diputacions provincials	B
Consells comarcals	B
Ajuntaments (el pla pilot pot incloure 10 ajuntaments[40])	AB
Entitats metropolitanes	C
Recerca	A
Universitats (el pla pilot pot incloure les 12 universitats catalanes[41])	A
Unitats R+D[42] (inclou els membres de la Xarxa d'Innovació Tecnològica; la Xarxa de Trampolins Tecnològics; i la Xarxa de Parcs Científics i Tecnològics de Catalunya, entre d'altres)	B
Humanitats (inclou entitats relacionades amb l'estudi de la filosofia i pensament, la teologia, la cultura, la literatura i la història)	C
Finances i treball	AB
Empreses (el pla pilot pot incloure les cambres de comerç i 10 empreses)	B
Col·legis professionals (el pla pilot pot incloure 10 col·legis professionals[43])	AB
Gremis, sindicats i patronals	C
Borsa de Barcelona	C
Vida associativa i Lleure	A
Associacions, agrupacions, fundacions, ONGs, etc. (el pla pilot pot incloure 10/20 associacions[44])	A

[40] En una selecció aleatòria: Badalona, Granollers, Terrassa, Figueres, Blanes, Balaguer, Sort, Àger, Reus, Tortosa,.

[41] Universitat de Barcelona (UB); Universitat Autònoma de Barcelona (UAB); Universitat Politècnica de Catalunya (UPC); Universitat Pompeu Fabra (UPF); Universitat de Lleida (UdL); Universitat de Girona (UdG); Universitat Rovira i Virgili (URV);Universitat Ramon Llull (URL); Universitat Oberta de Catalunya (UOC); Universitat de Vic (UVic); Universitat Internacional de Catalunya (UIC); Universitat Abat Oliba CEU (UAO).

[42] Directori disponible [consulta juliol 2005:] http://www10.gencat.net/dursi/ca/re/directori_r_d.htm

[43] En una selecció aleatòria: Col·legi d'Arquitectes de Catalunya; Col·legi de Fisioterapeutes de Catalunya; Col·legi de Periodistes de Catalunya; Col·legi d'Economistes de Catalunya; Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya; Col·legi Oficial de Químics de Catalunya; Il·lustre Col·legi Oficial de Veterinaris de Lleida; Col·legi Oficial d'Advocats de Barcelona; Col·legi Professional de Disseny Gràfic de Catalunya; Col·legi Oficial d'Aparelladors i Arquitectes Tècnics de Girona.

[44] En una selecció aleatòria: Associació Catalana d'Estacions d'Esquí i Activitats de Muntanya; Agrupació Astronòmica de Barcelona (ASTER); AGRUPANS (Agrupació per a Pares i Nens Superdotats de Catalunya); Asociación para la Recuperación de la Memoria Histórica; Associació Amics del Ferrocarril; Associació Conèixer Catalunya (ACCAT); Associació d'Empresàries i Executives; Associació Independent de Joves Empresaris de Catalunya (AIJEC); Associació de Mestres Rosa Sensat; Associació Unió Romaní; Òmnium Cultural; Pallassos sense Fronteres; Societat Catalana Cooperativa Limitada Abacus; Unió de Consumidors de Catalunya (UCC); Associació Catalana de Compositors; Joventuts

Institucions i espais culturals	B
Confessions religioses a Catalunya	B
Partits polítics amb representació parlamentària (el pla pilot pot incloure els 8 partits polítics[45])	A
Mitjans de comunicació[46]	AB
Premsa diària (inclou El 9 Nou; Avui; Diari de Barcelona; Diari de Girona; El Periódico; El Punt; Regió7; Segre; Sport; La Vanguardia, etc.)	B
Premsa periòdica en català (inclou L'Avenç; Cavall Fort; Digit-HUM; Monde Diplomatique (cat); Osona.com; Priorat Digital; El Temps; El Triangle; etc.)	A
Informatius en línia (inclou 3/24. Canal de notícies; Catalunya Informació en línia; e-notícies.com; Futur Català; TeleNotícies en línia; Vilaweb; etc.)	A
Agències de notícies (inclou Comunicació 21; En Joc; INTRA-ACN. Agència catalana de notícies; etc.)	B
Cadenes de televisió (inclou Televisió de Catalunya; Televisió Espanyola a Catalunya; i els membres de la Xarxa de Televisions Locals, etc.)	
Emissores de ràdio (inclou Cadena SER; Catalunya Ràdio; COMRàdio; Flaix FM; Ona Catalana; Onda Rambla; Rac1; etc.)	

Per tant, es pot realitzar una primera línia de treball, tenint en compte les diferències que existeixen entre els diversos tipus d'institucions que formen el grup preferent. En negreta, l'agent prioritari.

Generalitat de Catalunya i entitats associades	Ajuntaments	Universitats	Associacions	Mitjans de comunicació
President de la Generalitat Generalitat de Catalunya. Departament de la Presidència (inclou el Consell Català de l'Esport; l'Entitat Autònoma del Diari Oficial i de Publicacions de la Generalitat de Catalunya; l'Institut Català de la Dona; i el Patronat Català Pro-Europa) **Generalitat de Catalunya. Departament de Cultura** (inclou equipaments culturals consorciats: Biblioteca de Catalunya)				
		Generalitat de Catalunya. Departament d'Economia i Finances (inclou l'Institut Català de Finances; l'Institut d'Estadística de Catalunya) Generalitat de Catalunya. Departament d'Educació (inclou la Xarxa Telemàtica Educativa de Catalunya, XTEC) Generalitat de Catalunya. Departament d'Universitats, Recerca i Societat de la Informació		

Generalitat de Catalunya i entitats associades	Ajuntaments	Universitats	Associacions	Mitjans de comunicació
	Ajuntament de Badalona **Ajuntament de Granollers** **Ajuntament de Terrassa** **Ajuntament de Figueres** **Ajuntament de Blanes**			
	Ajuntament de Balaguer **Ajuntament de Sort** **Ajuntament d'Àger** **Ajuntament de Reus** **Ajuntament de Tortosa**			

Musicals de Catalunya; Casal Lambda; Associació Catalana de Premsa Comarcal; Associació d'Artistes Visuals de Catalunya.

[45] Ciutadans Pel Canvi (CPC); Convergència Democràtica de Catalunya (CDC);Esquerra Republicana de Catalunya (ERC); Esquerra Unida i Alternativa (EUIA); Iniciativa per Catalunya Verds (IC-V); Partit dels Socialistes de Catalunya (PSC); Partit Popular (PP);Unió Democràtica de Catalunya (UDC).

[46] Directori disponible [consulta juliol 2005:] http://cultura.gencat.net/mitjans/index.htm

Generalitat de Catalunya i entitats associades	Ajuntaments	Universitats	Associacions	Mitjans de comunicació

Universitat de Barcelona (UB)
Universitat Autònoma de Barcelona (UAB)
Universitat Politècnica de Catalunya (UPC)
Universitat Pompeu Fabra (UPF)
Universitat de Lleida (UdL)
Universitat de Girona (UdG)
Universitat Rovira i Virgili (URV)
Universitat Ramon Llull (URL)
Universitat Oberta de Catalunya (UOC)
Universitat de Vic (UVic)
Universitat Internacional de Catalunya (UIC)
Universitat Abat Oliba CEU (UAO)

Generalitat de Catalunya i entitats associades	Ajuntaments	Universitats	Associacions	Mitjans de comunicació

Col·legi d'Arquitectes de Catalunya
Col·legi de Fisioterapeutes de Catalunya
Col·legi de Periodistes de Catalunya
Col·legi d'Economistes de Catalunya
Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya
Col·legi Oficial de Químics de Catalunya
Il·lustre Col·legi Oficial de Veterinaris de Lleida
Il·lustre Col·legi Oficial d'Advocats de Barcelona
Col·legi Professional de Disseny Gràfic de Catalunya
Col·legi Oficial d'Aparelladors i Arquitectes Tècnics de Girona

Associació Catalana d'Estacions d'Esquí i Activitats de Muntanya
Agrupació Astronòmica de Barcelona (ASTER)
AGRUPANS (Agrupació per a Pares i Nens Superdotats de Catalunya)
Asociación para la Recuperación de la Memoria Histórica
Associació Amics del Ferrocarril
Associació Conèixer Catalunya (ACCAT)
Associació d'Empresàries i Executives
Associació Independent de Joves Empresaris de Catalunya (AIJEC)

Associació de Mestres Rosa Sensat
Associació Unió Romaní; Òmnium Cultural
Pallassos sense Fronteres
Societat Catalana Cooperativa Limitada Abacus
Unió de Consumidors de Catalunya (UCC)
Associació Catalana de Compositors
Joventuts Musicals de Catalunya
Casal Lambda
Associació Catalana de Premsa Comarcal
Associació d'Artistes Visuals de Catalunya

Ciutadans Pel Canvi (CPC)
Convergència Democràtica de Catalunya (CDC)
Esquerra Republicana de Catalunya (ERC)
Esquerra Unida i Alternativa (EUIA)
Iniciativa per Catalunya Verds (IC-V)
Partit dels Socialistes de Catalunya (PSC)
Partit Popular (PP)
Unió Democràtica de Catalunya (UDC)

Generalitat de Catalunya i entitats associades	Ajuntaments	Universitats	Associacions	Mitjans de comunicació

L'Avenç
Cavall Fort
Digit-HUM
Osona.com
El Triangle
Catalunya Informació en línia
e-notícies.com
Futur Català
TeleNotícies en línia
Vilaweb

Acords amb els agents

Seguidament, cal analitzar el tipus d'acord al que s'arribarà amb els agents implicats. Ja s'ha apuntat el triple objectiu de la Biblioteca de Catalunya, i per tant cal garantir a màxims:

- Que l'agent subministri sistemàticament el recurs digital quan es produeixin canvis significatius o amb la periodicitat que s'acordi si els canvis són diaris.
- Que l'agent faciliti la informació necessària en relació a la identificació del recurs que diposita, sigui per metadades o altres llenguatges que s'acordi.
- Que l'agent faciliti la informació necessària pel que fa al format del recurs que diposita, incloent llenguatges de programació, tipologia de fitxers no estàndards, etc.
- Que l'agent autoritzi a conservar el recurs dipositat, incloent la transformació que es pugui realitzar en el futur per l'ús de les tècniques necessàries, amb l'objectiu de preservar el contingut i la forma original del recurs.
- Que amb les limitacions temporals pertinents l'agent autoritzi a publicar la totalitat del recurs en obert i en línia o per defecte dins de terminals de consulta i connectats a la xarxa tancada i interna de la BC.
- L'agent podrà, si ho creu convenient, identificar el recurs vigent amb el segell "Padicat", que certifica de la institució la voluntat de contribuir a la preservació del patrimoni digital de Catalunya.

La Biblioteca de Catalunya, per la seva banda, ha de ser capaç de garantir a màxims:

- Facilitar una via senzilla i econòmica per realitzar el dipòsit amb la periodicitat que s'acordi.
- Facilitar el tractament professional de les dades relatives a la identificació del recurs que es diposita.
- Preservar en la mesura del possible la conservació en forma i contingut dels recursos digitals, emprant les eines i tècniques que tingui a l'abast per aquesta finalitat.
- Promocionar les entitats vinculades al Padicat, per mitjà de la presència de la URL del recurs vigent juntament amb els recursos dipositats, i altres formes que reforcin aquesta cooperació.

En tot cas, cal preveure que el tractament de les entitats susceptibles de ser contactades per al Padicat no serà igual en el cas d'institucions amb volum elevat de recursos digitals (administració, universitats) que en el cas d'institucions de baixa producció i periodicitat, que tanmateix formen part del panorama digital que ens ocupa.

Ens remetem finalment al calendari previst per als acords en la primera fase o pla pilot del Padicat (a partir de febrer de 2006) amb la necessitat de planificar curosament el sistema de contacte i formalització dels acords, així com de la publicitat que se'n derivi, i tanmateix de la importància d'establir contactes amb els segments decisoris de les grans corporacions (administració, universitats) a fi de facilitar la tasca pedagògica que suposarà l'entrada en vigor de la fase destinada a la captació i seguiment dels acords.

No es pot subestimar la inversió en hores de personal que suposa la necessària gestió dels acords amb els agents implicats en la producció digital catalana. Serveixi per reforçar aquesta idea el fet que la gestió dels acords del projecte danès suposa una inversió triennal (bàsicament en personal dedicat a aquesta tasca) que ronda els 434.607 €, o sigui un 44% del pressupost total.

3.4. Aspectes legals

El dipòsit legal

Si considerem que la legislació que afecta el Dipòsit Legal (1971) no contempla el patrimoni digital, la difusió del fons del Padicat és l'únic precepte limitat per la Llei de Propietat Intel·lectual.

Son objeto de depósito legal los escritos, estampas, imágenes y composiciones musicales, producidas en territorio nacional, en ejemplares múltiples, con fines de difusión, hechos por procedimientos mecánicos o químicos.

Comprenderá por tanto:

a. Libros, sea cualquiera la índole de su contenido y la forma de impresión y estén o no destinados a la venta.
b. Folletos, o sea escritos cuyo número de páginas sea mayor de cuatro y no exceda de 50, y con características semejantes a las señaladas en el párrafo anterior, incluyéndose en este concepto las separatas de artículos de revista que tengan la acotada extensión.
c. Hojas impresas con fines de difusión y que no constituyan propaganda esencialmente comercial.
d. Publicaciones periódicas (revistas y diarios).Partituras musicales.
e. Grabados: láminas sueltas, láminas de calendario, estampas, cromos, "chrismas", anuncios artísticos. Mapas y planos.
f. Carteles anunciadores de espectáculos, fiestas y demás actos públicos, tanto religiosos como profanos; anunciadores de artículos comerciales, siempre que lleven grabados artísticos; bandos y edictos.
g. Postales ilustradas.
h. Naipes.
i. "Slides" destinadas a difusión y venta.
j. Impresiones o grabaciones sonoras realizadas por cualquier procedimiento o sistema empleado en la actualidad o en el futuro.
k. Producciones cinematográficas, tanto de tipo argumental como documental, y "filmlets".

El text legal[47], de 1971, que regula les adquisicions de la Biblioteca de Catalunya per dipòsit legal no contempla la producció digital. En positiu, l'únic que contempla la regulació és la tipologia documental que afecta l'objectiu del Dipòsit legal, o sigui la "missió essencial de recollir tota la producció bibliogràfica nacional" (art. 5). Concretament, en el seu article 9:

De fet, l'objecte abstracte que centra la concepció de la missió del Dipòsit Legal (producció bibliogràfica) inclou la producció digital si es considera aquesta bibliogràfica, en el sentit biblioteconòmic del concepte, que prescindeix del suport material que contingui la informació.

Així, segons el tipus d'informació que contingui, és *bibliogràfic o cognitiu* aquell document de contingut cultural o cientificotècnic, reproduït en múltiples còpies, adreçat a un receptor anònim, el gran públic[48]. Malgrat tot, aquesta definició no pot incloure totes les produccions fruit dels llenguatges multimèdia de la publicació web, atès que pel mateix autor, l'altra gran família de documents, els *administratius o de gestió*, són els generats o rebuts per una persona o organització en l'exercici de l'activitat que li és pròpia. El problema de la definició acadèmica rau en què una web sovint conté unitats d'informació bibliogràfica i unitats d'informació administrativa.

Així, prescindint del suport i format del document, i ens centrem en el concepte "producció bibliogràfica", hem d'incloure la producció digital a banda de la que ja es recull a la Llei de 1971.

Però la realitat és més complexa: la situació d'*alegalitat* de la producció digital en relació al Dipòsit Legal ha facilitat, en tots els països del món --i a semblança del control bibliogràfic de la producció impresa, realitzat per les biblioteques nacionals segles més tard dels primers documents impresos— no s'hagi produït un dipòsit sistematitzat de la producció digital per part dels agents productors que són, més que mai, protagonistes de tot el cicle editorial (disseny+programació+redacció+publicació al servidor).

En tot cas, com indica Josep Vives[49], disposem de bons i suficients arguments per convèncer als nostres dipositants de la bondat dels repositoris, sense entrar en debats estèrils sobre la legalitat o no de preservar la producció digital catalana. A títol d'exemple: ni a Suïssa ni als Països Baixos existeixen lleis de dipòsit legal.

---

[47] "Orden de 30 de octubre de 1971, por la que se aprueba el Reglamento del Instituto Bibliográfico Hispánico", en *Boletín Oficial del Estado*, (18 nov 1971). En el capítol II es regula el Dipòsit Legal. L'ordre fou modificada parcialment per la "Orden de 20 de febrero de 1973 del Ministerio de Educación y Ciencia por la que se modifican algunos artículos del Reglamento del Instituto Bibliográfico Hispánico", en *Boletín Oficial del Estado* (3 mar 1973).

[48] Abadal, Ernest... [et al.]. *La Documentació a l'era de la informació*. Barcelona: Edicions de la Universitat Oberta de Catalunya, 1998. ISBN 84-95131-01-3

Llei de Propietat Intel·lectual

Més enllà de l'obsoleta regulació del dipòsit legal espanyol, un referent legal que ens afecta, pel fet de compilar i sobretot conservar i publicar en obert és la Llei de Propietat Intel·lectual[50] (LPI), per la qual l'autor d'una obra *decidirà si la seva obra ha de ser divulgada i en quina forma* (art. 14.1) o, per posar un altre exemple, *tota cessió [de drets] haurà de formalitzar-se per escrit* (art. 45).

Així, entendrem que els recursos digitals, del tipus que siguin, són objecte de propietat intel·lectual:

Són objecte de propietat intel·lectual totes les creacions originals literàries, artístiques o científiques expressades per qualsevol mitjà o suport, tangible o intangible, conegut actualment o que s'inventi en el futur. (art. 10).

No entrarem a valorar en profunditat com hem de considerar una "seu web" a efectes d'unitat d'aplicació de la legislació vigent. Sí citarem la definició[51] emprada habitualment pels membres del Laboratori d'Internet del CINDOC-CSIC:

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

A partir del fet que les interpretacions teòriques de la Llei són diverses[52], s'analitza a continuació la possible aplicació de la LPI a les fases més importants del projecte Padicat:

- En la captura dels recursos digitals que són d'accés obert a Internet (per exemple, una pàgina web) es produeix la localització del recurs, i la còpia a una unitat de memòria local.

- La localització del recurs (la visita a una URL) no està regulada. Per la pròpia naturalesa d'Internet, el fet d'accedir a un recurs que es una comunicació pública, difícilment la llei pot ser restrictiva.

---

[49] Vives, Josep. "Aspectos de propiedad intelectual en la creación y gestión de repositorios institucionales", en *El profesional de la información*. V. 14, n. 4 (jul-ago 2005). [consulta juliol 2005:] http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/267.pdf

[50] "Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el Texto Refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia", en *Boletín Oficial del Estado* (13 abr 1996).

[51] Interessant reflexió terminològica a: Pareja, Víctor Manuel; Ortega, José Luis; Prieto, José Antonio; Arroyo, Natalia; Aguillo, Isidro "Desarrollo y aplicación del concepto de sede web como unidad documental de análisis en Cibermetría", en *Jornadas Españolas de Documentación (9as: 2005 : Madrid)*. Madrid: Fesabid, 2005.

[52] Dos textos especialment rigorosos són: Vives, Josep. "Guia bàsica de propietat intel·lectual per a biblioteques", en *Ítem*, núm. 38 (2004), p. 103-152.; i Vives, Josep. "Aspectos de propiedad intelectual en la creación y gestión de repositorios institucionales", en *El profesional de la información*. V. 14, n. 4 (jul-ago 2005). [consulta juliol 2005:] http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/267.pdf

- La segona part de la fase de captura sí és susceptible de limitació per la Llei, si considerem "reproducció" a la còpia de la URL a la unitat local de memòria:

Correspon a l'autor l'exercici exclusiu dels drets d'explotació de la seva obra en qualsevol forma i, en especial, els drets de reproducció, distribució, comunicació pública i transformació, que no podran ser realitzades sense la seva autorització, excepte en els casos previstos a la present Llei. (art. 17)

- Aquest condicionant, però, no afecta la Biblioteca de Catalunya, que s'empara en la mateixa Llei:

Els titulars dels drets d'autor no podran oposar-se a les reproduccions de les obres, quan aquelles es realitzin sense finalitat lucrativa pels museus, biblioteques, fonoteques, filmoteques, hemeroteques o arxius, de titularitat pública o integrades en institucions de caràcter cultural o científic, i la reproducció es realitzi exclusivament per a fins de recerca. (art. 37.1)

- En l'emmagatzemament d'aquests recursos digitals (per exemple, en un disc dur o en un DVD) i les còpies d'aquests recursos per garantir la preservació.

- Com en el cas anterior, sí és susceptible de limitació per la Llei, si considerem "reproducció" a la còpia de la URL a la unitat local de memòria o altres suports:

Correspon a l'autor l'exercici exclusiu dels drets d'explotació de la seva obra en qualsevol forma i, en especial, els drets de reproducció, distribució, comunicació pública i transformació, que no podran ser realitzades sense la seva autorització, excepte en els casos previstos a la present Llei. (art. 17)

- I també com en el cas anterior, aquest condicionant no afecta la Biblioteca de Catalunya, que s'empara en la mateixa Llei, si la finalitat de la reproducció és d'investigació (no es contempla la "preservació" en l'excepció de la LPI):

Els titulars dels drets d'autor no podran oposar-se a les reproduccions de les obres, quan aquelles es realitzin sense finalitat lucrativa pels museus, biblioteques, fonoteques, filmoteques, hemeroteques o arxius, de titularitat pública o integrades en institucions de caràcter cultural o científic, i la reproducció es realitzi exclusivament per a fins de recerca. (art. 37.1)

- En la possible difusió del fons del Padicat a les dependències de la Biblioteca de Catalunya, o bé en línia i en obert, per exemple a través de la web de la BC.

- Queda afectada per tres dels preceptes legals que contempla la LPI, el de la mateixa definició del que és la difusió ("comunicació pública", en el llenguatge la Llei) i en el fet que cal una autorització prèvia:

S'entén per comunicació pública qualsevol acte pel qual una pluralitat de persones pot tenir accés a l'obra sense que prèviament se n'hagin distribuït exemplars a cadascuna (art. 20).

Correspon a l'autor l'exercici exclusiu dels drets d'explotació de la seva obra en qualsevol forma i, en especial, els drets de reproducció, distribució, comunicació pública i transformació, que no podran ser realitzades sense la seva autorització, excepte en els casos previstos a la present Llei. (art. 17)

Correspon a l'autor els següents drets irrenunciables i inalienables: Decidir si la seva obra ha de ser difosa i en quina forma. (art. 14.1)

- I és de dubtosa apreciació que hi ha diferències entre incorporar informació protegida per la llei en a les sales de la BC o a la web pública: l'article 20 no aprecia aquestes diferències, i el concepte de *pluralitat de persones* en l'activitat de *comunicació pública* pot cobrir tant si es tracta de tres com de tres-centes[53].

Esmentar finalment, i per a benefici d'aquesta institució, que el projecte de Llei[54] que hauria de modificar la LPI per adequar-la a l'entorn digital contempla que:

Bibliotecas y establecimientos análogos podrán poner a disposición de los investigadores obras que formen parte de sus fondos, siempre que la difusión se realice a través de terminales de consulta especializados y conectados a una red cerrada e interna.

Sense entrar a valorar la definició de "terminals de consulta especialitzats i connectats a una xarxa tancada i interna", podem preveure que la Biblioteca de Catalunya podria oferir les webs capturades a les seves dependències, com passa actualment a Suècia.

En tot cas, amb suport legal o sense, la voluntat de les entitats productores (universitats i centres de recerca, empreses i gremis, administració, col·legis professionals, associacions, partits polítics i sindicats, particulars) ha de veure el Padicat com l'oportunitat de ser els protagonistes de la futura recerca, així com de formar part del Patrimoni Digital de Catalunya, entès com un sistema útil per a la societat i les institucions que la formen.

L'experiència d'altres projectes

La situació que s'ha reflectit aquí és similar en la resta de territoris als que fan referència els projectes analitzats en l'informe pertinent[55].

[53] Vives, Josep. "Guia bàsica de propietat intel·lectual per a biblioteques", en *Ítem*, núm. 38 (2004), p. 103-152
[54] "Se remite a las Cortes Generales PROYECTO DE LEY por la que se modifica el texto refundido de la Ley de Propiedad Intelectual..." [en línia], en *Consejo de Ministros* (22 jul 2005). [consulta juliol 2005:] http://www.la-moncloa.es/web/asp/gob05.asp#PropiedadIntelectual

De l'anàlisi de les lleis del dipòsit legal dels diferents països es desprèn que habitualment la interpretació de la Llei marca la pauta: el concepte "document", en abstracte, és utilitzat per diverses biblioteques nacionals per compilar, processar i en menor mesura, difondre en obert, les webs capturades dels respectius territoris.

Es presenten els casos destacables en una breu panoràmica[56]:

- Països amb legislació (vigent, o en procés imminent d'aprovació) que afecta totes les formes d'edició digital (en suports físics i en línia):

    - Canadà

La *National Library Act* és de 1953[57] i ha tingut diverses reformes posteriors, essent la darrera de 1995[58].

Malgrat que el text legal inclou documents en vídeo, enregistraments sonors, CD-ROMs i microformes, la clau rau en la interpretació de "books" a la que afecta la Llei: s'hi considera qualsevol ítem que hagi estat "publicat".

En 1994-95 es produí l'*Electronic Publicactions Pilot Project* (EPPP) destinat a examinar el possible dipòsit de publicacions digitals en línia. Des de l'inici d'aquest pla pilot, la Biblioteca Nacional ha continuat compilant publicacions digitals en un dipòsit voluntari. S'emfatitza en publicacions no disponibles en altres formats.

    - Dinamarca

Traduïda a l'anglès com *Act on Copyright Deposit of Published Works*[59] i sancionada el gener de 1998 reemplaça la llei de 1927 i les esmenes de 1939 i 1991 (relatives a material no-llibre). El darrer canvi, de 2005, permet la Biblioteca nacional danesa capturar qualsevol web.

---

[55] *Estat de la qüestió arreu del món* (juliol 2005).
[56] Basat en un exhaustiu article a cura d'Amadeu Pons "El dipòsit legal dels recursos digitals: estat actual de la legislació en diversos països i projectes per endegar dipòsits nacionals de recursos digitals", en *Biblioteques digitals i dipòsits nacionals de recursos digitals* (Barcelona : 1999). Barcelona: Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, 1999, p. 25-54; i tanmateix en la web de la biblioteca nacional australiana: *Legal deposit* [en línia]. Canberra: National Library of Australia, 2003.[consulta juliol 2005:] http://www.nla.gov.au/padi/topics/67.html ; així com en el sintètic recull: Cordón, José Antonio. "El Depósito legal y los recursos digitales en línea" [en línia], en *Las bibliotecas nacionales del siglo XXI* (Valencia: 2005). [consulta agost 2005:] http://bv.gva.es/documentos/Ponencias/Cordon.pdf
[57] Canadà. *National Library Act* [consulta agost 2005:] http://laws.justice.gc.ca/en/N-12/index.html
[58] Canadà. *National Library Book Deposit Regulations* (1995). [consulta agost 2005:] http://laws.justice.gc.ca/en/N-12/SOR-95-199/index.html
[59] Dupont, Henrik. "Legal diposit in Denmark: the new law and electronic products" [en línia], en *LIBER Quarterly, the journal of European research libraries*, vol. 9 (1999), No 2. [consulta agost 2005:] http://liber-maps.kb.nl/articles/dupont11.htm

Com en el cas anterior, tots els materials "publicats" estan subjectes a la Llei, sense discriminació del tipus de producció tècnica, incloent-hi els recursos digitals en suports físics (CD-ROM...), i les publicacions en línia estàtiques o dinàmiques.

Un sistema de dipòsit en línia permet la biblioteca nacional gestionar el dipòsit voluntari, que només és consultable a la biblioteca i no pot ser copiat pels usuaris, però sí imprès.

- Itàlia

La nova *Norme relative al deposito legale dei documenti di interesse culturale destinati all'uso pubblico*[60] (2004) substitueix el text anterior, de 1939, i la revisió de 1945.

Inclou genèricament, a banda del materials tradicionals, els documents difosos en suport informàtic i els difosos per mitjà de xarxa informàtica. En la pràctica, la institució dipositària, la biblioteca nacional, és responsable de la captura de la web italiana[61].

- Noruega

La *Pliktavleveringsloven* (*Norwegian Legal Deposit Act*[62], en anglès) és vigent des de juliol 1990, malgrat ser aprovada el 1989.

Inclou tots els documents que s'hagin fet públics, en un sentit ampli, i deixa així la porta oberta a possibles nous suports.

- Nova Zelanda

El capítol 4 de la *National Library of New Zealand Te Puna Matauranga o Aotearoa Act 2003*[63] cobreix el dipòsit legal, reemplaçant així el text legal anterior, de 1965.

Afecta els documents públics impresos o electrònics (suport físic o en línia), i permet la biblioteca nacional recopilar els documents físics o copiar documents a Internet, per mitjà de la consulta als productors digitals.

---

[60] Itàlia. *Norme relative al deposito legale dei documenti di interesse culturale destinati all'uso pubblico* [consulta agost 2005:] http://www.parlamento.it/parlam/leggi/04106l.htm

[61] Comentaris a la Llei a la web de l'AIB: *Gruppo di studio sulle biblioteche digitali* [consulta agost 2005:] http://www.aib.it/aib/commiss/bdigit/deplegdig.htm

[62] Noruega. *The legal deposit of generally available documents* [consulta agost 2005:] http://www.pliktavlevering.no/html/legal_deposit.html

- Regne Unit

El text de la *Legal Deposit Libraries Act 2003*[64] va substituir la secció 15 de la *Copyright Act* de 1911 i les posteriors modificacions de 1956 i 1998.

D'una anàlisi somer es desprèn que el nou text legal obre el dipòsit legal britànic als materials no impresos[65], i a cobrir així CD-ROMs, publicacions seriades en línia i seus web.

- Sud-àfrica

La *Legal Deposit Act*[66], data de 1997, però és vigent des de 1998.

L'amplia definició de "document" (traducció lliure: *qualsevol objecte el qual s'ha previst per conservar o transmetre informació en format textual, gràfic, visual, sonor o altre format intel·ligible, mitjançant qualsevol mitjà...)* i la interpretació del "mitjà" (traducció lliure: *qualsevol tipus d'enregistraments o transmissió d'informació realitzat per la subsegüent lectura, audició o vista*) permet que la Llei s'apliqui als documents electrònics en suport físic i tanmateix en línia.

- Suècia

La llei vigent és de 1993[67] (revisada en 1995), i té els seus precedents en l'any 1661

EL text inclou els recursos digitals en suport físic. Nogensmenys, el decret[68] de 8 de maig de 2002 autoritza la biblioteca nacional sueca a compilar les seus web sueques i permetre el públic consultar-les en les seves dependències. Es pot imprimir però no enregistrar les webs de la col·lecció.

- Països amb legislació que afecta l'edició digital en suport físic:
  - Alemanya

La llei data de 1969, i es refereix a tot document imprès, sonor i audiovisual. En la pràctica, s'hi inclou els suports físics amb informació digital.

---

[63] Nova Zelanda. *National Library of New Zealand Te Puna Matauranga o Aotearoa Act 2003.* [consulta agost 2005:] http://www.legislation.govt.nz/libraries/contents/om_isapi.dll?clientID=827261728&infobase=pal_statutes.nfo&jump=a2003-019&softpage=DOC

[64] Regne Unit. *Legal Deposit Libraries Act 2003* [consulta agost 2005:] http://www.opsi.gov.uk/acts/acts2003/20030028.htm

[65] Alvestrand, Viveka. "UK legal deposit law moves into digital age"en *Information world review*, 197 (ec 2003), p. 1. Una notícia de 2004 a la premsa generalista es feia ressò dels esforços en la preservació del patrimoni digital amb un titular prou impactant: Hudson, Rebeca. "Saviours of the lost archives", en *Sunday Times*, (2 may 2004), p. 12.

[66] Sud-àfrica. *Legal Deposit Act* (1997). [consulta agost 2005:] http://www.dac.gov.za/legislation%5Fpolicies/acts/0rb%7Elegal%20deposit%20act%2054%20of%201997.doc

[67] Suècia. *Legal Deposit Act*. [consulta agost 2005:] http://www.kb.se/Ple/eng/act.htm

[68] Suècia. *Decree 2002.287 regarding the treatment of personal information in the Royal Library's digital project of cultural heritage*. [consulta agost 2005:] http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm

- Àustria

La *Novelle zum Mediengesetz*[69] (2000), substitueix la de 1981, referida només a material imprès. El nou text inclou els recursos digitals en suport físic.

- Estats Units d'Amèrica

El dipòsit legal als Estats Units està regulat per la secció 407 de la *Copyright Act of 1976*[70], i inclou els materials en format tradicionals, amb diverses addendes (fitxers informàtics, des de 1988, i a partir d'aquest, les cintes magnètiques, els CD-ROMs, etc.)

- França

La darrera versió de la llei data de 1992[71]. La legislació s'aplica, explícitament, als documents sobre un suport físic, cosa que exclou qualsevol document publicat en línia.

- Suïssa

No compta amb legislació federal (sí el cantó de Friburg). A la pràctica, la biblioteca nacional té la responsabilitat dels documents en qualsevol suport, incloent els documents electrònics fixos.

- Països amb projectes de dipòsit digital nacional sense legislació específica:
  - Austràlia

La llei vigent és de 1968[72]. Inclou material imprès, pel·lícules i discos. En alguns estats autònoms s'hi ha afegit publicacions electròniques, i la biblioteca nacional australiana ha endegat un programa intern de dipòsit voluntari de publicacions electròniques.

- Finlàndia

La llei és de 1980. Inclou publicacions impreses, els enregistraments sonors i de vídeo. Una llei posterior, de 1984, es refereix a pel·lícules cinematogràfiques. Hi ha una sèrie de propostes posteriors per incloure les publicacions digitals en format físic.

---

[69] Àustria. *Novelle zum Mediengesetz*. [consulta agost 2005:] http://www.onb.ac.at/about/lza/mg-novelle-2000-bgbl.pdf
[70] Estats Units d'Amèrica. *Copyright Act of 1976*. [consulta agost 2005:] http://www.copyright.gov/title17/chapter04.pdf
[71] Game, Valerie. "Depot legal et droit a l'ere du numerique" [en línia], en *Propietá intelettuale e nuove tecnologie in biblioteca* (7 mai 2004). [consulta agost 2005:] http://www.comune.milano.it/webcity/documenti.nsf/0/7c721efe134a3ddbc1256e97003f8fca/$FILE/Game.pdf
[72] Austràlia. *Copyright Act 1968*. [consulta agost 2005:] http://www.austlii.edu.au/au/legis/cth/consol_act/ca1968133

- Japó

La *National Diet Library Law* data de 1948 (revisió de 1949), incloent els documents en format tradicional. És previst incloure CD-ROMs en la futura revisió[73].

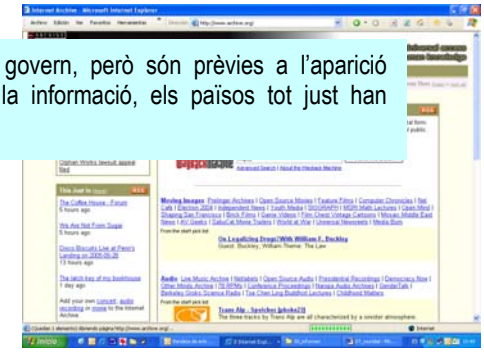
- Països Baixos

No hi ha legislació sobre dipòsit legal, però es recull pràcticament tot el material imprès a partir d'acords voluntaris entre la biblioteca nacional i els editors. Entre els acords globals més importants, destacar el *Depot van Nederlandse Elecktronische Publicaties* (DNEP[74]), impulsat el 1995, i origen de l'e-Depot.

- El cas Internet Archive

Protegint el nostre dret a conèixer

Alguns països tenen Lleis que garanteixen l'accés públic als documents del govern, però són prèvies a l'aparició d'Internet. Malgrat que Internet ha incrementat generalment l'accés públic a la informació, els països tot just han començat a canviat aquestes lleis per reflectir l'actual *societat Internet*...



Co licat als Estats Units, la informació er re inclou text, àudio, imatge en moviment, i programari, així com pàgines web arxivades de tot el món. El recurs conté en accés obert en línia 35 milions de seus web, des de 1996 fins a l'actualitat, i cada dos mesos es realitza una captura massiva que afecta 4.000 milions de pàgines web. Per què no té problemes legals un ens que captura i posa en obert les seus web capturades? Algunes respostes, en traducció lliure, que podem trobar a la web de l'Internet Archive:

Exercint el nostre "dret a recordar"

Sense biblioteques de paper, seria difícil exercir el nostre "dret a recordar" la nostra història política o la responsabilitat que tenen els governs. Amb molts dels assumptes públics passant del paper a l'entorn digital, les biblioteques d'Internet estan esdevenint essencials en mantenir aquest dret. Imagineu, per exemple, com es ressentiria la cobertura periodística d'una campanya electoral si els periodistes tinguessin accés limitat a les declaracions prèvies dels candidats...

Com puc retirar la meva web de l'Internet Archive?

L'Internet Archive no està interessat en preservar o oferir accés a seus web o altres documents d'Internet de persones que no volen els seus materials en la col·lecció. Afegint un simple arxiu robots.txt en el vostre servidor, podeu excloure la vostra seu web de les captures, així com excloure qualsevol pàgina història de l'Internet Archive...

[73] Japó. *Legal Deposit System Council*. [consulta agost 2005:] http://www.ndl.go.jp/en/aboutus/deposit_council_book.html
[74] Països Baixos. *Depot van Nederlandse Elecktronische Publicaties.* [consulta agost 2005:] http://www.kb.nl/dnp/e-depot/dm/geschiedenis-en.html
[75] Kimpton, Michele. "Saving the web for future generations", *Archiving web resources: international conference* (Canberra: nov 2004). [en línia]. Canberra: National Library of Australia, 2005. [consulta juny 2005:] http://www.nla.gov.au/webarchiving/MicheleKimpton.ppt

3.5. Conclusions sobre el context a Catalunya: els recursos existents, agents implicats, aspectes legals

- El Padicat no té precedents a Catalunya ni a Espanya. Però sí es té experiència, especialment per mitjà del CBUC, a la cooperació entre agències, com són exemples les TDX o el RACO.

- La BC haurà d'arribar a acords amb diversos agents per assegurar compilar amb regularitat i exhaustivitat les seves webs, i poder oferir en obert aquest patrimoni digital. A major espectre temàtic, institucional, territorial, major serà la repercussió del projecte a Catalunya.

- No es pot subestimar l'alta inversió en hores de personal que suposa la gestió dels acords amb els agents.

- Si considerem que el Dipòsit Legal no inclou la producció digital, l'únic escull legal és la difusió, la "comunicació pública" de les webs capturades.

4. Disseny del sistema d'informació: abast, captura, organització i accés als recursos[76]

4.1. Introducció

El juny de 2005, la Biblioteca de Catalunya va iniciar el projecte PADICAT (Patrimoni Digital de Catalunya), que té per objectiu dissenyar i produir un sistema que permeti la BC compilar, processar i donar accés permanent a la producció digital catalana.

En alguns països reben el nom de "dipòsits digitals nacionals" els projectes similars, essent els més coneguts el gegant Internet Archive (http://www.archive.org), l'australià Pandora (http://pandora.nla.gov.au/index.html) o el suec Kulturarw3 (http://www.kb.se/kw3/ENG/).

D'acord amb la tendència generalitzada a les biblioteques nacionals, el model de dipòsit que persegueix la Biblioteca de Catalunya és el sistema híbrid, consistent a:

- Compilar massivament els recursos digitals publicats en obert a Internet
- Impulsar el dipòsit 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.

En documents anteriors[77] al present informe s'ha detallat les tasques realitzades des de l'inici del projecte, i concretament s'ha analitzat l'estat de la qüestió arreu del mon; el context a Catalunya, per mitjà dels recursos existents, els agents implicats i els aspectes legals que poden condicionar el sistema; així com diversos treballs dirigits a planificar (calendari, recursos humans, costos) el projecte en les fases de producció i explotació, previstes per al període 2006-2008.

És objectiu del present informe presentar el sistema d'informació en la seva fase de disseny, tot considerant els aspecte relatius a aquest sistema, com són l'abast de la col·lecció, i els processos de captura, organització i accés als recursos.

---

[76] El present capítol, complementat amb la resta dels que es recullen en el present informe, ha donat lloc a la presentació i acceptació de la ponència (inèdita): Llueca, Ciro. "El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya", *Jornades Catalanes d'Informació i Documentació* (10es : 2006: Barcelona). Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 2006.

[77] Llueca, Ciro. *Memòria del plantejament del projecte PADICAT.* Barcelona: Biblioteca de Catalunya, 2005, que conté els informes "Estat de la qüestió arreu del món", "Context a Catalunya: recursos existents, agents implicats, aspectes legals", "Planificació a curs i mig termini i fases d'execució", "Recursos humans necessaris per executar el projecte: perfils i tasques", i "Estudi de costos vinculats a la fase de producció". Tots s'hi inclouen en el present informe global.

L'informe es completarà, a mida que la fase avanci, amb el test i les proves que quedaran documentades a l'informe corresponent, i que en bona mesura poden afectar a l'abast del projecte, així com a a la compilació, emmagatzematge i difusió de la col·lecció.

4.2. Abast del PADICAT

Cal en primer lloc una definició el més clara possible de quins és la tipologia de recursos tecnològics publicats a Internet, i quines les temàtiques, que són susceptibles de considerar part del Patrimoni Digital de Catalunya.

Com a norma bàsica, entenem "Patrimoni Digital" com la informació electrònica publicada a Internet, en obert o no, independentment del format en què es presenta aquesta informació. Entendrem "de Catalunya" en el sentit que tradicionalment ha tingut la bibliografia nacional de Catalunya en què es basa la política de la Biblioteca de Catalunya: tot allò produït a Catalunya o que tracti sobre Catalunya. Es definirà el concepte de "comunitat web de Catalunya", que servirà de marc referencial del que considerem "de Catalunya".

4.2.1. 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 emprada en relació a les unitats d'informació que representa cada seu web, però sí citarem la definició[78] emprada habitualment pels membres del Laboratori d'Internet del CINDOC-CSIC, que servirà per definir el que genèricament entenem com a web:

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.

[78] Interessant reflexió terminològica a: Pareja, Víctor Manuel; Ortega, José Luis; Prieto, José Antonio; Arroyo, Natalia; Aguillo, Isidro "Desarrollo y aplicación del concepto de sede web como unidad documental de análisis en Cibermetría", en *Jornadas Españolas de Documentación (9as: 2005 : Madrid)*. Madrid: Fesabid, 2005.

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

- 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 genèrica, així com el tractament que caldrà seguir el procés del que coneixem tradicionalment com *parts components*[79], 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 treballs[80] de l'International Internet Preservation Consortium, 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.

4.2.2. Abast temàtic

Com han recollit els teòrics[81], Internet està dissenyada per trencar les barreres geogràfiques i fer la informació accessible universalment. 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.

---

[79] 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.

[80] Marill, Jennifer [et al.] *Web harvesting survey*. [en línia]. International Internet Preservation Consortium (version 1, jul 2004) http://netpreserve.org/publications/iipc-r-001.pdf i Boyko, Andrew [et al.] *Test bed taxonomy for crawler*. [en línia]. International Internet Preservation Consortium (version 1, jul 2004) http://netpreserve.org/publications/iipc-r-002.pdf

[81] Daniel Gomes, Mário J. Silva, "Characterizing a National Community Web", [en línia] en *ACM Transactions on Internet Technology*, volume 5, number 3, August 2005. http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf [consulta: desembre 2005]

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

- El cas australià[82] 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[83], que inclou:
    - 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[84], 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[85] 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:

---

[82] "What is the scope of the Archive? Do you have selection guidelines?" [en línia], en Pandora (http://pandora.nla.gov.au/panfaqs.html#scope) [consulta: desembre 2005]
[83] "Defining Swedish web pages and finding them?", [en línia] en Kulturarw3 (http://www.kb.se/kw3/ENG/Description.htm) [consulta: desembre 2005]
[84] *Act in Legal Deposit of Published Material: traslation of Act No. 1439 of 22 December 2004: unauthorized version*. [en línia] http://www.bs.dk/content.aspx?itemguid=%7b332484E6-A5B1-4CEE-B953-059843182050 [consulta: desembre 2005]

- 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 i en especial de l'últim cas, definim la informació electrònica susceptible de formar part del PADICAT el grup de documents que contenen informació relativa a Catalunya o d'interès majoritari de la gent catalana, aspecte que conceptualment queda ja recollit a l'article 7 de la Llei de biblioteques de 1981[86]:

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. 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:

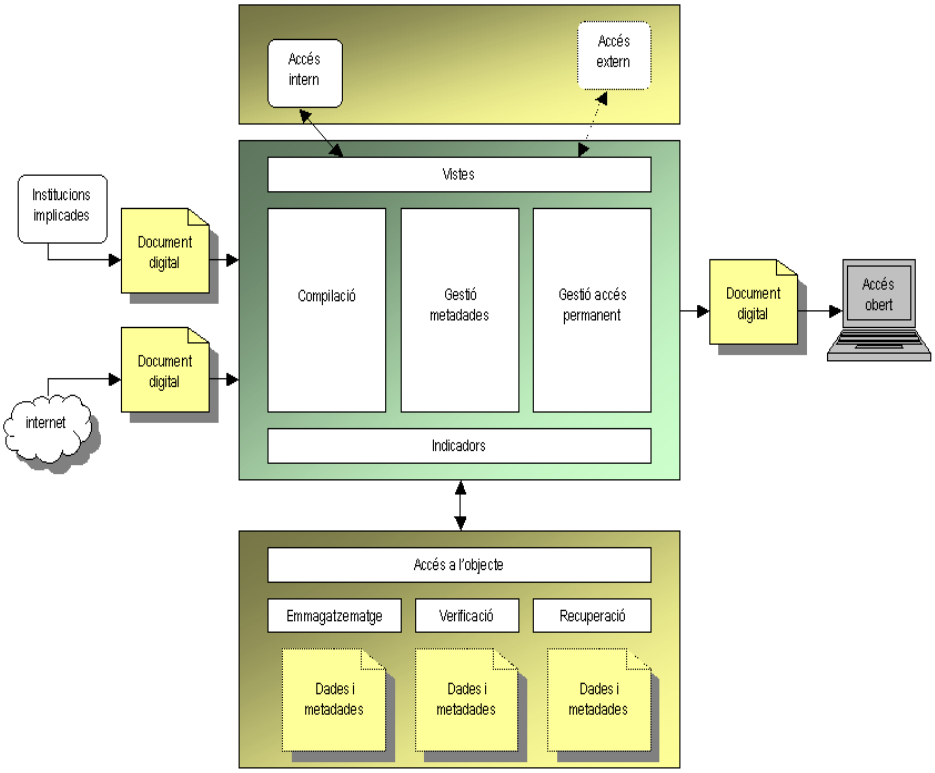
- Webs sota domini .CAT[87],
- Webs ubicades a servidors de Catalunya[88], o
- Webs sota dominis geogràfics (.ES, .COM, .NET, .ORG, etc[89].) en llengua catalana[90], o
- Webs que no compleixin els requisits anteriors, però relacionades temàticament amb Catalunya[91]

---

[85] Daniel Gomes, Mário J. Silva, "Characterizing a National Community Web", [en línia] en *ACM Transactions on Internet Technology*, volume 5, number 3, August 2005. http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf [consulta: desembre 2005]
[86] "Llei de biblioteques de Catalunya, de 24 d'abril de 1981", *Diari Oficial de la Generalitat de Catalunya*, núm. 123 (29 abr 1981).
[87] L'associació puntCAT (http://www.puntcat.org) informa 04/12/2005 que la posada en funcionament del domini és prevista a gener de 2006.
[88] Es definirà els paràmetres en el capítol dedicat a la compilació, així com en l'informe dedicat a la fase de proves del projecte.
[89] A partir del treball de Ricardo Baeza-Yates, Carlos Castillo i Vicente López a "Characteristics of the Web of Spain" [en línia], en *Cybermetrics*, Vol. 9 (2005): Issue 1. Paper 3. http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html, 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.
[90] Es definirà els paràmetres en el capítol dedicat a la compilació, així com en l'informe dedicat a la fase de proves del projecte.
[91] A partir de directoris, com Dmoz, Google, Yahoo, etc.

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.

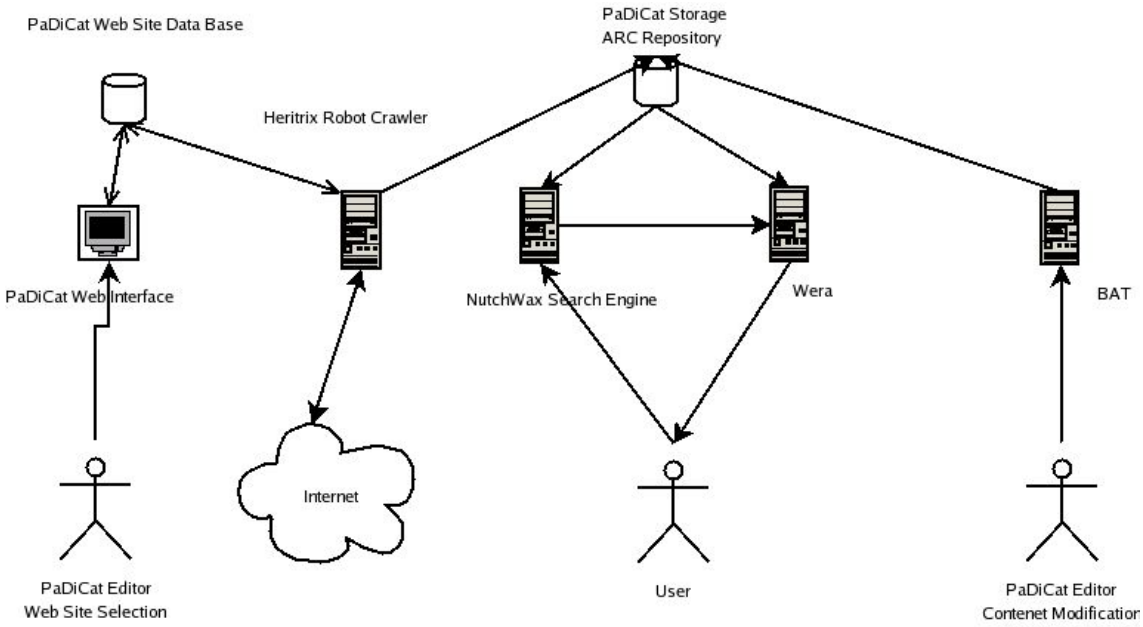
Es proposa en el gràfic un model[92] amb la descripció somera dels aspectes relacionats amb les parts components, que d'altra banda només difereixen en qüestions puntuals del cicle documental clàssic de les biblioteques i serveis d'informació (adquisició + tractament + difusió).



4.3. Sistema d'informació

El sistema, tal com apareix al gràfic[93], es basa en els diversos mòduls del programari Heritrix[94], completats amb una base de dades MySQL que agrupa els registres relatius als recursos web que formen la col·lecció.

[92] Basat en Dulabahn, Beth. "The National Digital Information Infrastructure and Preservation Program (NDIIP): future directions and relevance to other countries", *Archiving web resources: international conference* (Canberra: nov 2004). [en línia]. Canberra: National Library of Australia, 2005. [consulta abril 2005:] http://www.nla.gov.au/webarchiving/program.htm
[93] Obra de Leandro Stasi, enginyer informàtic del projecte PADICAT.
[94] Característiques del programa en l'informe dedicat al test del projecte.



En base a la imatge precedent identifiquem les següents parts:

- L'editor [PaDiCat Editor], que alimenta la base de dades i en revisa continguts amb l'administrador BAT
- La interfície de l'editor [PaDiCat Web Interface], que facilita l'accés i organització de la col·lecció
- La base de dades [PaDiCat Web Site Data Base], que inclou els registres dels recursos que formen la col·lecció
- El robot [Heritrix Robot Crawler], dedicat a compilar els recursos d'Internet en base a la col·lecció
- Internet, on es troben publicats els recursos web
- El dipòsit [PaDiCat Storage ARC Repository], on es troben els recursos web capturats
- L'indexador [Nutchwax Search Engine], que indexa el dipòsit i en possibilita la cerca i recuperació
- La interfície de consulta [Wera], que possibilita l'accés, cerca i recuperació dels recursos
- L'usuari extern [user], que accedeix a la col·lecció via la interfície de consulta

- El sistema d'administració [BAT], que permet l'editor revisar-ne els continguts

Conceptualment identifiquem les parts claus del procés en la captura dels recursos, l'organització dels recursos, i l'accés permanent als recursos.

4.4. Captura dels recursos

La compilació dels recursos s'orienta en les línies principals[95] que s'han apuntat en la introducció, o sigui:

- Compilar massivament els recursos digitals publicats en obert a Internet
  - Automatitzada o integralment[96]
  - Manualment
  - 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[97] i maquinari[98] que s'empren en el test d'aquesta fase 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, com els problemes relacionats amb el temps de captura, el rebuig per les instruccions del

[95] D'acord també amb Birte Christensen-Dalsgaard [et al.] *Final report for the pilot project "netarkivet.dk"* [en línia].
[96] *Snapshot* en llengua anglesa, que podríem traduir com *foto fixa*.
[97] Heritrix (versió 1.6, desembre de 2005) és un compilador (crawler, en llengua anglesa) de codi obert desenvolupat per l'Internet Archive amb les National Nordic Libraries. Més informació a: http://crawler.archive.org
[98] Dos PCs amb processador Intel Pentium IV 3.2GH, 2GB de RAM i Disc Dur d'1,2 TB (3x400GB)

fitxer Robot.txt[99], l'anàlisi dels *logs* per l'accés denegat, etc. En la fase pilot del projecte s'abordaran aquests conflictes i les vies de resolució.

A continuació s'analitzarà les diverses línies de captura tenint en compte el seu procediment i abast conceptual.

4.5. Compilació massiva dels recursos digitals publicats en obert a Internet

El sistema preveu una base de dades [PaDiCat Web Site Data Base] que integra el llistat de recursos web que formen la col·lecció. Aquesta base de dades es fonamenta en els recursos que procedeixen del robot de cerca [Heritrix Robot Crawler] i que són supervisats i organitzats pel personal dedicat a la gestió i organització [PaDiCat Editor].

En tot cas, es preveu una tasca automatitzada d'identificació dels recursos web, en paral·lel a la implementació manual dels recursos que per selecció s'hi vulguin afegir.

4.5.1. Captura automatitzada

	Domini .CAT	Servidor a Catalunya	Llengua catalana	Relació temàtica
Automatitzada	X	X	X	
Acció	URL sota domini .CAT	IP a territori Catalunya	Llengua catalana en el document	X

La fase pilot (2006) del projecte estarà dedicada pràcticament en la seva integritat a compilar per captura automatitzada els recursos web de Catalunya.

En base al que s'ha descrit, els tres eixos de treball són:

Webs sota domini .CAT

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

[99] 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 escropolosament. Netarkivet (Dinamarca) no ho respecta mai.

catalana a Internet, des de l'inici del seu funcionament. La posada generalitzada en funcionament del domini és prevista a gener de 2006[100].

Webs ubicades a servidors[101] de Catalunya

Està pendent de definir el conjunt d'accions dirigides a obtenir els codis IP[102] dels servidors instal·lats a territori de Catalunya.

L'organisme encarregat de l'assignació d'IP és l'IANA (Internet Assigned Numbers Authority, http://www.iana.org), i la seva secció europea és RIPE (Réseaux IP Européens, http://www.ripe.net). RIPE ofereix[103] 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[104] 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[105] 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 ESNIC (Network Information Center para España, http://www.esnic.es) és una altra fórmula per obtenir llistats dels dominis els organismes dels quals tinguin seu a Catalunya[106] sota domini .ES, que a data de la present redacció són un total de 45.374[107].

Webs en llengua catalana

Està pendent de definir el conjunt d'accions dirigides a identificar òptimament la llengua catalana a les webs de qualsevol domini.

---

[100] El 21/12/2005 s'anuncia que la web www.domini.cat és la primera amb el domini .CAT.
[101] Devem aquestes informacions en bona mesura 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.
[102] De fet, conèixer les IP que operen a ubicacions de Catalunya té limitacions perquè en molts casos hi ha IP que hostatgen més d'una web.
[103] En una cerca aleatòria dirigida als organismes amb la paraula "Barcelona" trobem 100 resultats (possiblement un màxim predefinit) que concorden a aquest criteri.
[104] http://www.iana.org/assignments/ipv4-address-space
[105] Empreses com Ip2location (http://www.ip2location.biz), Phpcontrol (http://www.phpcontrol.com) o Maxmind (http://www.maxmind.com).
[106] Iniciats els tràmits de consulta amb ESNIC a desembre de 2005.
[107] La xifra inclou els dominis .ES (37.354), .COM.ES (6.013), .ORG.ES (1.025), .NOM.ES (936), .EDU.ES (30), i .GOB.ES (16) amb titulars residents a Madrid.

La detecció de la llengua catalana és possiblement un dels factors determinats en la compilació de webs per al projecte PADICAT. I és que, més enllà dels dominis[108] geogràfics on aquestes s'hostatgin) el sol fet de contenir la llengua catalana és indicador de pertinença a la bibliografia catalana.

En aquest sentit es va sol·licitar suport a dues institucions que podien aportar la seva expertesa al projecte: l'Institut d'Estudis Catalans[109] i la Universitat Politècnica de Catalunya[110].

A partir de les indicacions d'aquestes institucions, inferim[111] que pel que fa a freqüències de paraules:

- Determinades fórmules gramaticals són pròpies, i sovint exclusives, de la llengua catalana: l'ela geminada [l·l]; determinats pronoms febles a continuació de formes verbals [...ar-ne, ...re-se]; article masculí plural i pronom personal [els].

- Determinades combinacions de paraules o paraules en solitari poden definir molt probablement un contingut amb llengua catalana: *haver* i *es*; *amb* i *com*; *seu* o *seva* o *seua* o *seus* o *seves* o *seues*; *més* i *tot*; *molt*; *això*; *perquè*; *sense* i *ho*; *nosaltres*; etc.

4.5.2. Captura manual

	Domini .CAT	Servidor a Catalunya	Llengua catalana	Relació temàtica
Manual				X
Acció				Entrada manual dels URL susceptibles de captura

A partir de la revisió de les web integrades a la col·lecció per la via automatitzada cal preveure l'ampliació o revisió manual cap a nous recursos que puguin formar part, en una primera instància, d'altres webs.

En aquest sentit, el sistema permet la identificació de recursos (via la seva URL) que calgui incloure manualment al robot de captura, juntament amb la resta de dades que en l'organització de la col·lecció s'acordi adoptar.

---

[108] A partir del treball de Ricardo Baeza-Yates, Carlos Castillo i Vicente López a "Characteristics of the Web of Spain" [en línia], en *Cybermetrics*, Vol. 9 (2005): Issue 1. Paper 3. http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html, 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.

[109] Devem a Joan Soler i Bou, responsable del Diccionari de Freqüències de l'Institut d'Estudis Catalans part de les aportacions d'aquest capítol.

[110] Contactada la UPC en desembre de 2005, estem a l'espera de resposta.

[111] A falta de tancar la fase de test no s'ha realitzat un experiment en aquest sentit, amb una base prou significativa de recursos.

4.5.3. Compilació per dipòsit voluntari sistemàtic de la producció web dels agents implicats a Catalunya

	Domini .CAT	Servidor a Catalunya	Llengua catalana	Relació temàtica
Dipòsit voluntari				X
Acció				Identificació del productor[112], anàlisi de la periodicitat del dipòsit, contacte i acord, seguiment del dipòsit

El dipòsit voluntari de la producció web dels diversos agents a Catalunya ha de ser facilitada pel sistema en base a una plataforma eficaç que permeti els interessants assenyalar (al robot) o dipositar per les vies (FTP, etc.) que s'acordaran quan les webs siguin susceptibles de ser integrades a la col·lecció.

La plataforma, així com la interfície de treball, té producció prevista en l'any 2006.

4.5.4. Compilació focalitzada dels recursos digitals relacionats amb determinats esdeveniments de la vida pública catalana.

	Domini .CAT	Servidor a Catalunya	Llengua catalana	Relació temàtica
Esdeveniments				X
Acció				Identificar esdeveniment[113], Identificar mitjans, planificació periodicitats, seguiment i captura.

Al fil de la tendència generalitzada a les biblioteques nacionals amb projectes de dipòsit digital nacional, és previst iniciar una línia de recerca en base a la captura focalitzada i exhaustiva de tot el que es publiqui digitalment a Catalunya, en relació a un fet que considerem clau per a entendre la societat catalana.

[112] Un llistat inicial de productors susceptibles d'arribar a acords amb la Biblioteca de Catalunya és a l'informe Llueca, Ciro. *Memòria del plantejament del projecte PADICAT: context a Catalunya: recursos existents, agents implicats, aspectes legals*. Barcelona: Biblioteca de Catalunya, 2005

[113] Una possible fita és el referèndum sobre el nou estatut de Catalunya, previst per a la tardor de 2006.

En aquest sentit, la tasca manual de selecció i seguiment dels recursos web que hi estiguin dedicats té la recompensa és la immediatesa de la col·lecció creada en relació a l'esdeveniment focalitzat.

L'acció té producció prevista en l'any 2006.

4.6. 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.

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.

4.7. Identificació permanent

Com van explicar a bastament Muxach i Lopo[114] la descripció de recursos en la xarxa en basa en l'URI (Uniform Resource Identifer = 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.

Amb un exemple diferent al de Muxach i Lopo:

- Amb l'URL, http://www.bnc.es/bc/projectes.php#padicat consta de dues parts: <http> i <www.bnc.es/bc/projectes.php#padicat>.

- Amb l'URN, http://www.bnc.es/bc/projectes.php#padicat tindríem que <www.bnc.es> és l'URN que ens indica que ens trobem a les pàgines web de la Biblioteca de Catalunya que poden ser variables i diverses, però que estaran relacionades amb la Biblioteca de Catalunya.

- Amb l'URC, http://www.bnc.es/bc/projectes.php#padicat ens permetrà saber que "títol=pàgina principal de la BC; autor=Biblioteca de Catalunya, etc."

El cert a data d'avui, i com ja apuntaven aquests autors el 1999, és que la majoria de col·leccions continuen emprant l'URL en referir-se a la localització del document digital[115]. Malgrat la realitat, i en base als treballs[116] del CBUC i les previsions terminològiques que entenem vigents, es decideix la denominació URI (Uniform Resource Identifer = Identificador Uniforme de Recurs) per a cadascuna de les versions compilades dels recursos web digitals.

4.8. 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.).

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[117].

En tot cas, i com explica Gail M. Hodge[118], una vegada l'arxiu ha compilat el document digital, és necessari identificar-lo i catalogar-lo. 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. Cal analitzar els aspectes relatius a com s'ha

---

[114] Muxach, Santi i Ana Lopo. "Metadades a peu pla", en *Ítem*, núm. 24 (1999), p. 99-134. Disponible en línia:
[115] Hodge
[116] 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.
[117] Muxach, Santi; Lopo, Ana. "Metadades a peu pla", *Ítem*, núm. 24 (1999), p. 99-134. http://www.cobdc.org/cgi-bin/intranet/itemdoc.pl?page=num24/smuxach.pdf
[118] Hodge, Gail M. "Best practices for digital archiving: an information life cycle approach", en *D-LIB Magazine* [en línia], v. 6, núm. 1 (jan 2000). http://www.dlib.org/dlib/january00/01hodge.html

creat la metadada, els estàndards i les normes que s'han emprat, el nivell d'aplicació de les metadades, i on estan emmagatzemades aquestes.

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[119] a partir del model Dublin Core. Aquell document ha servit de base[120] a reduir per fer la present 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 (bibliografia nacional o no, etc.) al qual es refereixi.

Metadada	Especificacions	Nivell
contributor.author	*Persona, organització o servei responsable de la creació del contingut del document* Doneu la forma tal com consta en l'índex d'autors del CCUC En cas que en el CCUC trobeu més d'una forma o no en trobeu cap, escolliu aquella que s'adeqüi més a les vostres necessitats o la que aparegui en el propi document Si hi ha més d'un responsable en la creació del contingut del document repetiu l'element tantes vegades com sigui necessari. Doneu-los seguint l'ordre que apareix en el document Si el responsable és una entitat (organització o servei), doneu-lo en la primera casella Per a responsabilitats sobre el contingut del document, altres que l'autoria, useu l'element "contributor"	mínim
contributor	*Persona, organització o servei responsable de fer contribucions al contingut del document* Useu aquest element per donar contribucions altres que l'autoria (ex.: traductors, il·lustradors, prologuistes) i les que s'informen per plantilla Doneu la forma tal com consta en l'índex d'autors del CCUC En cas que en el CCUC trobeu més d'una forma o no en trobeu cap, escolliu aquella que s'adeqüi més a les vostres necessitats o la que aparegui en el propi document Si hi ha més d'un responsable de fer contribucions repetiu l'element tantes vegades com sigui necessari. Doneu-los seguint l'ordre que apareix en el document Si el responsable de fer contribucions és una entitat (organització o servei), doneu-lo en la primera casella En cas de dubte sobre el tipus de responsabilitat useu l'element "contributor.author"	avançat
title	*Títol donat al document* Nota: si el document té més d'un títol (abreviat, en una altra llengua, etc.) seleccioneu l'opció corresponent en la primera pantalla del formulari. Doneu el títol principal i, si és el cas el subtítol, en aquest element. Doneu altres títols en l'element "title.alternative" Independentment de la tipografia usada en el document, doneu el títol en minúscula (excepte inicials) i amb accents (si n'hi ha) *Els articles són ignorats per defecte a l'hora de l'ordenació, per tant no cal que els posposeu ni elimineu*	mínim
title.alternative	*Altre títol donat al document* Nota: si el document té més d'un títol (abreviat, en una altra llengua, etc.) seleccioneu l'opció corresponent en la primera pantalla del formulari	avançat

[119] 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.

[120] Devem aquesta informació a les aportacions i recomanacions de Francesca Navarro de la unitat de Digitalització de la BC i de Ida Conesa del Servei de Normalització Lingüística.

	Independentment de la tipografia usada en el document, doneu el títol en minúscula (excepte inicials) i amb accents (si n'hi ha) Si hi ha més d'un títol alternatiu repetiu l'element tantes vegades com sigui necessari *Els articles són ignorats per defecte a l'hora de l'ordenació, per tant no cal que els posposeu ni elimineu*	
date.created	Data de creació del contingut intel·lectual del document *En cas que no consti cap data en el document, doneu obligatòriament un any aproximat. Si el mes i el dia no el coneixeu no cal que en doneu cap d'aproximat*	avançat
publisher	*Entitat responsable de la publicació i/o distribució del document* Nota: si el document ha estat publicat i/o distribuït anteriorment seleccioneu l'opció corresponent en la primera pantalla del formulari Doneu la forma tal com consta en l'índex d'autors del CCUC En cas que en el CCUC trobeu més d'una forma o no en trobeu cap, escolliu aquella que s'adeqüi més a les vostres necessitats o la que aparegui en el propi document	avançat
language.iso	Llengua del contingut del document Trieu la llengua del llistat desplegable Si hi ha més d'una llengua repetiu l'element tantes vegades com sigui necessari	avançat
description.abstract	Breu resum del document Doneu aquest resum en la llengua del document Si la llengua del document no és el català, doneu-lo opcionalment també en aquesta llengua (cada resum en una casella diferent)	avançat
contributor	*Persona, organització o servei responsable de fer contribucions al contingut del document* Es dóna la universitat i el departament o grup de recerca, institut, etc. a la qual pertany la col·lecció del RECERCAT de la que forma part el document Es dóna aquesta entitat tal com consta en l'índex d'autors del CCUC Si us cal afegir dades, modificar la forma que trobeu en la plantilla o bé afegir una altra contribució, doneu-la també tal com consta en l'índex d'autors del CCUC. En cas que trobeu més d'una forma o no en trobeu cap, escolliu aquella que s'adeqüi més a les vostres necessitats o la que aparegui en el propi document	avançat
rights.uri	Condicions d'ús i reproducció del document Es dóna la llicència Creative Commons "Reconeixement–NoComercial-SenseObraDerivada"	avançat
identifier.uri	*URI que s'assigna a cada document i que és única i irrepetible*	per defecte
description.provenance	Història de la creació i modificacions posteriors del document	per defecte
format.extent	*Extensió del document en bytes*	per defecte

4.9. 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.

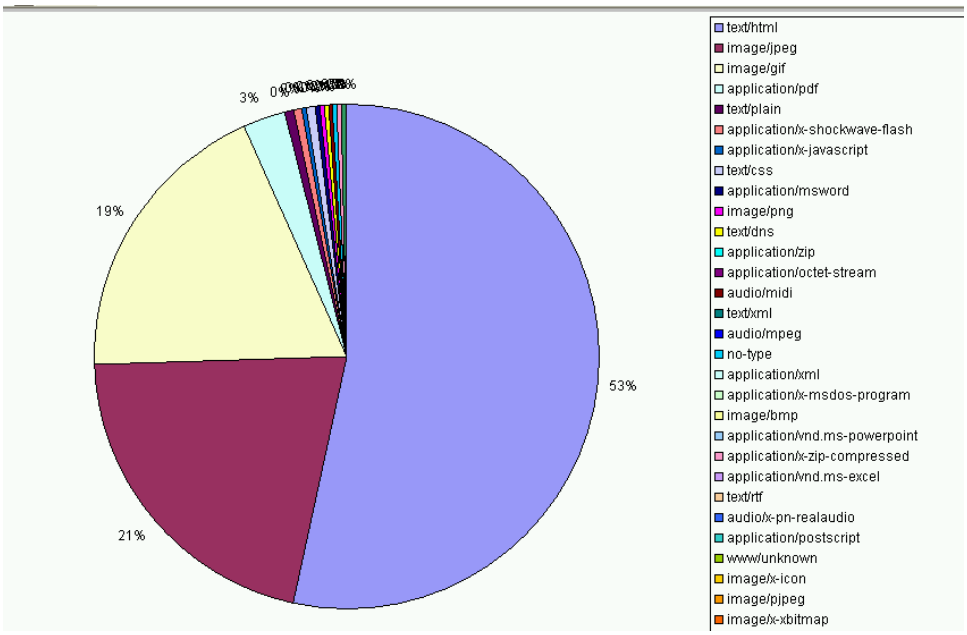
El programari emprat en el test permet avançar que els arxius s'emmagatzemen comprimits amb l'extensió estàndard .ARC.

És previst un capacitat necessària total de 10 TB en el període de producció del projecte (2006-2008).

4.10 Preservació

La preservació és l'aspecte de la gestió de la col·lecció que preserva el contingut i l'aparença (el *look and feel*) del document digital.

Les estratègies més habituals de preservació[121] 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.



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.

En un experiment realitzat per cobrir aquest detall de l'informe, 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%).

[121] Ayre, Catherine; Muir, Adrienne. "The right to preserve: the rights issues of digital preservation", *D-Lib magazine* [en línia], vol. 10, num. 3 (mar 2004). [consulta abril 2005:] http://www.dlib.org/dlib/march04/ayre/03ayre.html

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

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	application/x-msdos-program
1.912	0,064	image/png	317	0,139	image/bmp

El principal problema, en conseqüència, estarà relacionat amb els formats o aplicacions que són efímers, 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: és el que *s'anomena 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.

4.11. Vida útil del dipòsit

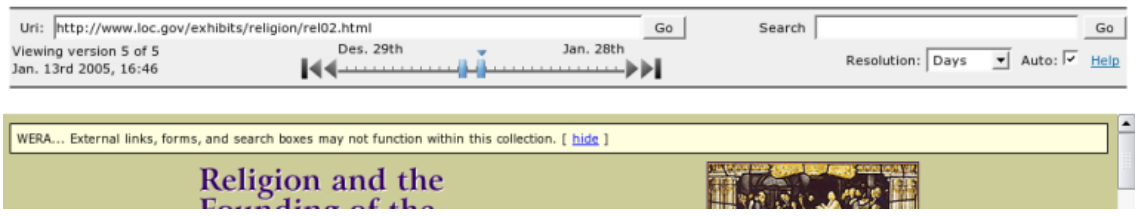
Recentment[122] s'ha desenvolupat el concepte, aplicat a projectes com el que ens ocupa, del MTTF (Mean Time to Failure), per calcular el temps de vida útil abans que es produeixi una falla irreversible que comenci a deteriorar parts del dipòsit.

En el cas danès aquest període s'estipula actualment en 144 anys.

4.12. 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 altre cas, 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 Heritrix, el Nutchwax Search Engine, que indexa els recursos abans de la compressió i en garanteix la cerca i resposta per mitjà de la interfície de consulta Wera (gràfic).



Wera ofereix un sistema de cerca i consulta basat en les opcions que l'administrador determini, i que per defecte són la cerca per URL, per text lliure, i la barra de navegació per les diferents versions del recurs.

[122] Christensen, Niels, "Preserving the bits of the Danish Internet", en Internet Web Archive Workshop (Viena, 2005), http://www.iwaw.net/05/christensen.pdf. [Consulta: desembre 2005]

5. Maquinari i programari necessari: plataformes i opcions (inclòs testeig de programari)

5.1. Introducció

El test es va realitzar durant els mesos de novembre i desembre de 2005, i gener de 2006.

L'empresa adjudicatària del servei, AUSEBA, S.A., va posar a disposició del projecte l'analista informàtic Leandro Stasi.

L'objectiu del servei era, tal com consta als plecs de la convocatòria:

> L'objecte és la contractació d'un servei d'implementació, testeig i avaluació de prototips del projecte PADICAT (Patrimoni Digital Català), d'acord amb la descripció del punt 2 [tasques].
>
> El projecte PADICAT (Patrimoni Digital Català) té com a objectiu crear un sistema d'informació basat en les tecnologies i les telecomunicacions, que permeti la creació i alimentació constant d'un repositori virtual dels recursos electrònics catalans en línia, és a dir del patrimoni digital català a Internet.
>
> La Biblioteca de Catalunya realitzarà fins a finals de l'any 2005 la implementació i testeig de prototips ja existents per a la captació, emmagatzemament, organització i accessibilitat de recursos electrònics en línia com a part del primer pas del projecte que ha de permetre a la BC disposar d'un projecte escrit que reculli detalladament les actuacions i processos tècnics i tecnològics implicats, el sistema informàtic, la plataforma i tecnologia necessària, així com d'un prototip testejat i avaluat.

Concretament, la descripció de les tasques incloïa:

- Determinar junt amb el cap del projecte PADICAT de la Biblioteca de Catalunya les possibilitats tecnològiques de recollida de recursos electrònics catalans en línia:
  - Dominis/recursos internet de l'entorn català sobre els quals es realitzarà la recollida de recursos electrònics.
  - Determinar les eines i mètodes per a obtenir automàticament els dominis/recursos de l'entorn català susceptibles de ser incorporats en la recollida de recursos.
  - Determinar les limitacions del model automàtic en l'obtenció de recursos electrònics de l'entorn català.
  - Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari de captació de recursos.
  - Determinar els volums de recollida òptims per als testeigs.
  - Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari d'indexació de recursos.
  - Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari de cerca i visualització de la base de dades generada a partir de la recollida i indexació dels recursos electrònics d'àmbit català.
- Recollir tota la informació resultant dels testeigs:
  - Objectius assolits i problemes detectats quan a l'obtenció automàtica de dominis/recursos en l'entorn català. Limitacions.
  - Volum de dades generat per a cada recollida.
  - Temps dedicat a cada recollida segons el nombre de urls recollit i segons els maquinari i les limitacions utilitzades en el testeig.
  - Anàlisi valoratiu dels testeigs del programari de recollida, del programari d'indexació i del programari de cerca i visualització de la base de dades de recursos electrònics d'àmbit català: costos en temps i maquinari, càlcul de requeriments mínims segons els objectius. Problemes i limitacions observades.
  - Adaptacions de programari realitzades.
  - Requeriments necessaris a nivell de telecomunicacions.
- Instal·lar i adaptar el sistema operatiu Linux a les màquines de proves.

- Instal·lar i configurar el programari base: servidor web (Apache 2), servidor d'aplicacions (Tomcat 5), etc.
- Instal·lar el programari inclòs com a prototips a estudi.
- Adaptar i parametritzar el programari segons les necessitats del projecte PADICAT.
- Testejar, avaluar i readaptar el programari fins assolir els objectius del projecte PADICAT.
- Documentar tots els passos realitzats amb indicació de:
  - Programari instal·lat: versions, pedaços, errors, limitacions.
  - Parametrització realitzada per a cada testeig.
  - Adaptacions realitzades per a cada testeig.
  - Objectius a assolir per a cada testeig.
  - Resultats obtinguts per a cada testeig.
  - Altres que es creguin necessaris en el moment de realitzar les proves.

  La documentació s'ha de lliurar regularment (com a màxim setmanalment) als responsables del projecte a la BC.
- Fer ús de les eines de consulta i contacte necessaris per a mantenir-se informat/da de les novetats de caràcter tecnològic que puguin sorgir en el període comprés en el contracte.

5.2. Valoració global del servei

Es presenta un quadre indicatiu del grau de compliment de les tasques del projecte per part del servei contractat, que estimem en un 83% respecte als objectius previstos.

L'ordenació dels epígrafs controlats varia en relació al plec de la convocatòria per adequar-los al procés natural d'implantació del servei.

Tasca	*Grau compliment*	*Observacions*
Instal·lar i adaptar el sistema operatiu Linux a les màquines de proves.	100%	La instal·lació del sistema operatiu Linux s'ha dut a terme, amb el suport de l'ATI de la BC, en el maquinari previst a tal efecte.
Instal·lar i configurar el programari base: servidor web (Apache 2), servidor d'aplicacions (Tomcat 5), etc.	100%	La instal·lació del programari base s'ha dut a terme, amb el suport de l'ATI de la BC, en el maquinari previst a tal efecte.
Instal·lar el programari inclòs com a prototips a estudi.	100%	La instal·lació dels mòduls del programari Heritrix (Heritrix, BAT, NutchWax, Wera) s'ha dut a terme en el maquinari previst a tal efecte.
Adaptar i parametritzar el programari segons les necessitats del projecte PADICAT.	80%	S'ha adaptat i parametritzat el programari segons les instruccions rebudes. Les fases posteriors hauran de completar aquesta parametrització del sistema.
Testejar, avaluar i readaptar el programari fins assolir els objectius del projecte PADICAT.	80%	S'ha testat i readaptat el programari segons les instruccions rebudes. Les fases posteriors hauran de completar aquesta parametrització del sistema.
Determinar junt amb el cap del projecte PADICAT de la Biblioteca de Catalunya les possibilitats tecnològiques de recollida de recursos electrònics catalans en línia.	70%	El test ha servit per determinar els dominis/recursos d'internet de Catalunya sobre els quals es realitzarà la recollida (60%); les eines i mètodes per obtenir automàticament els recursos a capturar (70%); les limitacions del model automàtic en l'obtenció dels recursos (80%); el nombre i metodologia per al test de captació (80%); volums òptims de recollida (80%); nombre i metodologia per al test d'indexació (70%); i nombre i metodologia per al test de cerca i visualització (70%).
Recollir tota la informació resultant dels testeigs.	90%	S'ha comprovat els objectius de captura i els problemes i limitacions dels recursos (90%); el volum de dades generat per cada recollida (80%); el temps dedicat a cada recollida (80%); l'anàlisi valoratiu del test del programari de recollida, indexació i cerca i visualització (70%); S'han realitzat les adaptacions de programari necessàries per a cadascuna de les limitacions

		observades (90%); s'ha testat els requeriments necessaris a venill de telecomunicacions (90%).
Documentar tots els passos realitzats amb indicació de...	40%	El temps dedicat al desenvolupament del programari i test del mateix no ha produït, en paral·lel, una documentació àmplia sobre aquests processos. El coordinador del projecte ha optat per acceptar aquesta via, tenint en compte les limitacions temporals de l'encàrrec.
Fer ús de les eines de consulta i contacte necessaris per a mantenir-se informat/da de les novetats de caràcter tecnològic que puguin sorgir en el període comprés en el contracte.	90%	L'analista s'ha incorporat a la llista [archive-crawler@yahoogroups.com] i hi ha participat activament proposant els dubtes del projecte, rebent instruccions de la resta d'analistes de biblioteques nacionals i projectes similars, per tenir les versions actualitzades i les aplicacions concretes de cada cas. No s'ha contactat de manera personalitzada amb cap analista en concret.
Valoració global	83%	

5.3. Instal·lar i adaptar el sistema operatiu Linux a les màquines de proves.

La instal·lació del sistema operatiu Linux s'ha dut a terme, amb el suport de l'ATI de la BC, en el maquinari previst a tal efecte.

5.4. Instal·lar i configurar el programari base: servidor web (Apache 2), servidor d'aplicacions (Tomcat 5), etc.

La instal·lació del programari base s'ha dut a terme, amb el suport de l'ATI de la BC, en el maquinari previst a tal efecte.

5.5. Instal·lar el programari inclòs com a prototips a estudi.

En l'anàlisi del programari existent en els projectes similars, i especialment arrel de la visita professional a les biblioteques nacionals de Suècia i Dinamarca, la trobada amb el responsable de la branca europea de l'Internet Archive, Julien Masanès, i tanmateix per l'assistència a l'International Web Archiving Workshop[123] (Viena, setembre 2005), es poden extraure les següents conclusions:

- A partir de 1996 diverses biblioteques nacionals han desenvolupat els programaris *ad hoc* per a projectes com el que ens ocupa, els més destacables són el Combine[124] (Suècia), Nordic Web Index[125] (biblioteques escandinaves), i PANDAS[126] (Austràlia). Tots els programes són de programari obert i es poden consultar, i en molts casos provar, a partir de les versions existents en línia.

[123] De periodicitat anual, la informació relativa a l'IWAW es troba disponible a: http://www.iwaw.net
[124] http://combine.it.lth.se/
[125] http://www.lib.helsinki.fi/finelib/kopenhamn/hakala2/nwinwa.ppt
[126][126] http://pandora.nla.gov.au/pandas.html

- Esporàdicament, i especialment en els darrers anys, s'han realitzat altres experiments amb programari propi per a projectes similars[127] al PADICAT o de dipòsit institucional[128].

- A partir de 1996 Internet Archive va desenvolupar un paquet de programes[129], sota el paraigua del que es coneix com Heritrix, al qual s'han anat sumant la resta de projectes, especialment les biblioteques nacionals escandinaves (Nordic National Libraries).

- En l'actualitat, la majoria dels projectes de dipòsit digital en funcionament (no en fase de proves, etc.) utilitzen total o parcialment[130] la plataforma Heritrix.

Entenem que l'ús d'Heritrix està generalitzat a la majoria de projectes, i això en base a una sèrie de característiques que n'afavoreixen la ràpida expansió i la consolidació:

- Eficàcia provada, Internet Archive, l'arxiu web gegant, utilitza des de 1996 les diverses versions d'aquest programari. En els darrers anys la major part dels projectes de les biblioteques nacionals s'han incorporat al desenvolupament i aplicació del programa en totes o determinades fases dels seus sistemes.

- Programari en codi obert consolidat, gràcies a una comunitat d'usuaris elevada en nombre de persones, que procedeixen de biblioteques nacionals que tenen confiats els seus projectes en aquest programari. El fet que el programari sigui en obert implica, a més, complir amb les recomanacions del DURSI en matèria d'ús de programari obert en les administracions públiques, i el mateix fet comporta, paral·lelament, una inversió mínima en despesa directa de compra de llicències o manteniment del sistema.

- Alta parametrització del seu funcionament, Heritrix dona resposta a les demandes de diverses vies de treball (sistema integral, selectiu, captura per domini, per IP, per cerca en text lliure, etc.), d'acord amb la diversitat dels projectes existents.

L'ús del programari lliure, a més, com indica Barragán[131], des del punt de vista filosòfic, convergeix en diferents punts amb l'esperit de servei en la difusió i preservació de la informació que és competència dels professionals de l'àmbit de la biblioteconomia i la documentació, si bé la bibliografia existent en les

[127] Grècia 2003, Eslovènia 2004, Portugal 2004.
[128] Linden, Jim [et al.]. *Technology watch report: the large-scale archival storage of digital objects*. London: The British Library, 2005. http://www.dpconline.org/docs/dpctw04-03.pdf
[129] Heritrix, BAT, NutchWax, i Wera.
[130] Austràlia i el Regne Unit empren PANDA, però han incorporat sistemàticament Heritrix en el seu funcionament. Suècia usa Combine, però dedica les seves captures més delicades (diaris, etc.) a Heritrix.
[131] Barragán, Cristina. "Programari lliure: Introducció i estat de la qüestió per als professionals de la informació i la documentació". *Bibliodoc 2004*. Barcelona: COBDC, 2005. p. 59-76.

publicacions especialitzades se centra majoritàriament en casos pràctics de desenvolupament i d'implementació de programes lliures en biblioteques.

Com s'ha esmentat, però, quan parlem de la plataforma Heritrix ens estem referint a tres programes diferents que interactuen per crear un sistema integral en codi obert de captura, gestió i recuperació dels dipòsits digitals: Heritrix, BAT, NutchWax, i Wera.

Heritrix, *el capturador*

Heritrix[132] és un robot de captura (*crawler*) de codi obert desenvolupat en les darreres versions per l'Internet Archive i les Nordic National Libraries, a partir de les millores que es proposen per part de la comunitat d'usuaris al voltant de la *Crawler Discussion List.*

Es tracta d'un programari que navega per Internet en base a unes coordinades prèvies. D'acord amb una sèrie de criteris configurats per l'administrador del sistema, Heritrix emmagatzema amb compressió ARC, o no, les webs que ha visitat, afegint-hi nous URL a la captura.

BAT, *el gestor d'arxius*

BAT[133] (*Bnf Arc Tools*) va ser desenvolupat i és mantingut per la Bibliothèque Nationale de France.

És un conjunt de mòduls capaços d'administrar i realitzar modificacions en els arxius del tipus ARC, DAT i CDX.

NutchWax Search Engine, *el cercador en el propi sistema*

NutchWax[134] (*Nutch Web Archive eXtensions*) és un motor de cerca de codi obert desenvolupat per la fundació Apache.

Té dues funcions bàsiques: d'una banda, crear índexs de les dades emmagatzemades; d'altra banda, permet la cerca en les dades del dipòsit, com ho faria un motor de cerca d'Internet.

Wera, *la interfície de consulta*

Wera[135] (*Web Archive Access*) va ser desenvolupat inicialment per la Nasjonalbiblioteket (Noruega), i actualment s'hi ha afegit l'Internet Archive.

---

[132] http://crawler.archive.org. Un article fonamental al respecte és: Mohr, Gordon [et al.]. "An introduction to Heritrix: an open source archival quality web crawler", a *International Web Archiving Workshop* (4th : 2004). http://www.iwaw.net/04/proceedings.php?f=Mohr
[133] http://crawler.archive.org/cgi-bin/wiki.pl?BnfArcTools
[134] http://archive-access.sourceforge.net/projects/nutch/

De fet és una interfície de cerca i visualització dels resultats que serveix per navegar, amb el motor de cerca NutchWax, dins d'un dipòsit de seus web.

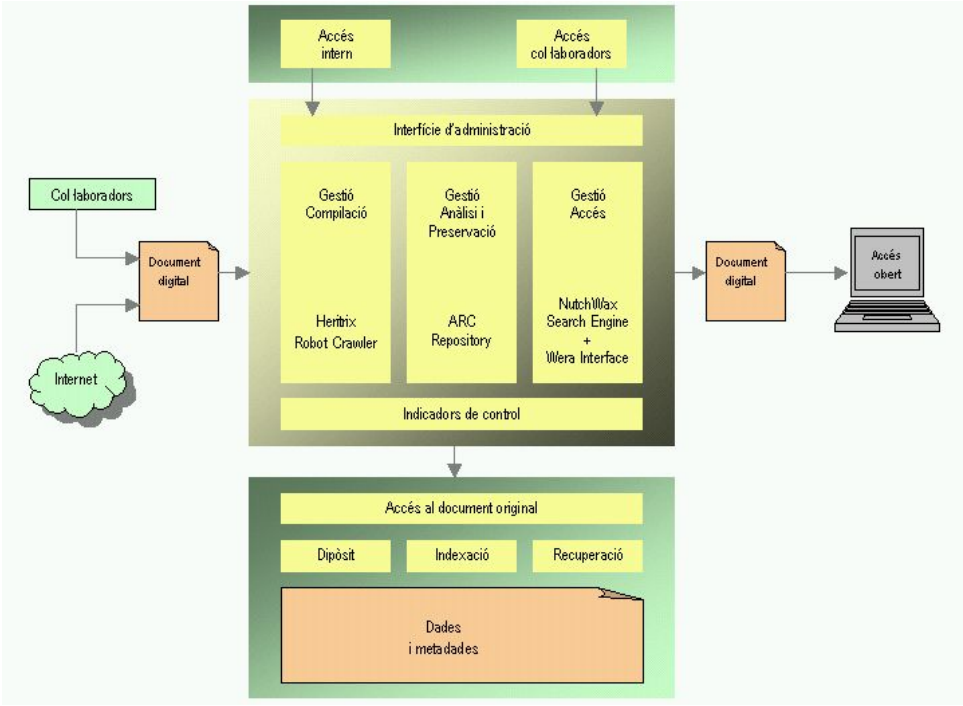
5.6. Adaptar i parametritzar el programari segons les necessitats del projecte PADICAT.

Com s'ha vist, el sistema híbrid que caracteritza el PADICAT ha de permetre la Biblioteca de Catalunya tres accions simultànies:

- Compilar automàticament els recursos que es defineixin
- Arribar a acords amb institucions per afavorir el dipòsit voluntari de les seves pàgines web.
- Focalitzar la captura en determinats esdeveniments claus de la societat catalana.

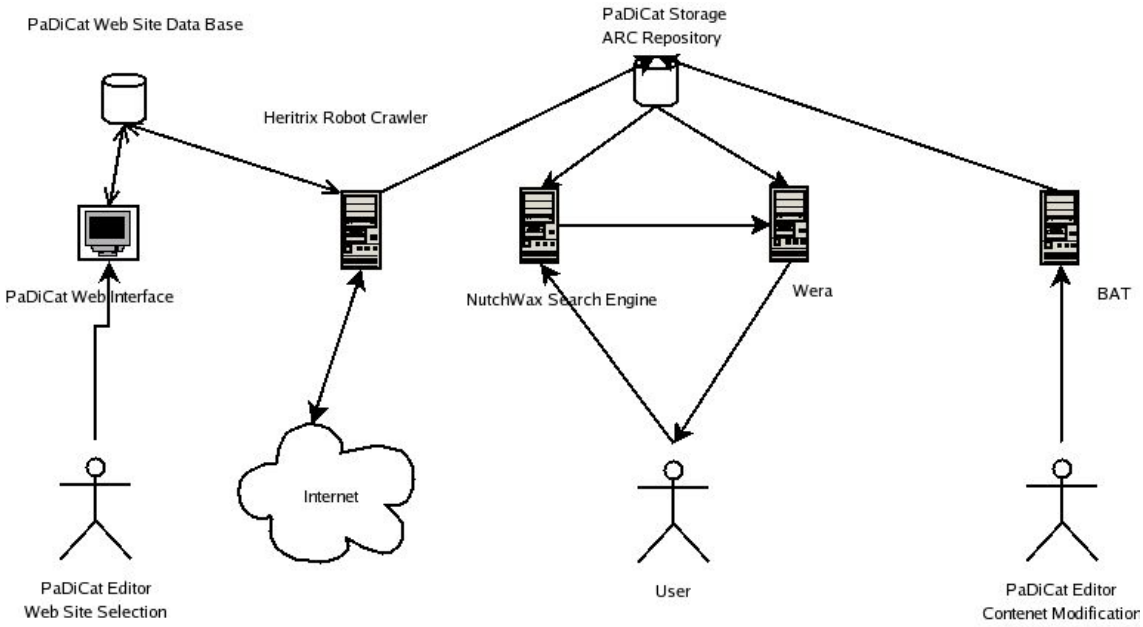
Es proposa en el gràfic un model conceptual[136] amb la descripció somera dels aspectes relacionats amb les parts components. Només difereix en qüestions puntuals del cicle documental clàssic de les biblioteques i serveis d'informació.

El sistema conceptual aplicat al programari del projecte, tal com apareix al gràfic[137] següent, es basa en els diversos mòduls del programari Heritrix, completats amb una base de dades MySQL que agrupa els registres relatius als recursos web que formen la col·lecció.



[135] http://archive-access.sourceforge.net/projects/wera/

[136] 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]



En base a la imatge precedent identifiquem les següents parts:

- L'editor [PaDiCat Editor], que alimenta la base de dades i en revisa continguts amb l'administrador BAT
- La interfície de l'editor [PaDiCat Web Interface], que facilita l'accés i organització del fons
- La base de dades [PaDiCat Web Site Data Base], que inclou els registres dels recursos que formen la col·lecció
- El robot [Heritrix Robot Crawler], dedicat a compilar els recursos d'Internet en base a la col·lecció
- Internet, on es troben publicats els recursos web
- El dipòsit [PaDiCat Storage ARC Repository], on es troben els recursos web capturats
- L'indexador [Nutchwax Search Engine], que indexa el dipòsit i en possibilita la cerca i recuperació
- La interfície de consulta [Wera], que possibilita l'accés, cerca i recuperació dels recursos
- L'usuari extern [user], que accedeix a la col·lecció via la interfície de consulta
- El sistema d'administració [BAT], que permet l'editor revisar-ne els continguts

[137] Obra de Leandro Stasi, enginyer informàtic del projecte PADICAT.

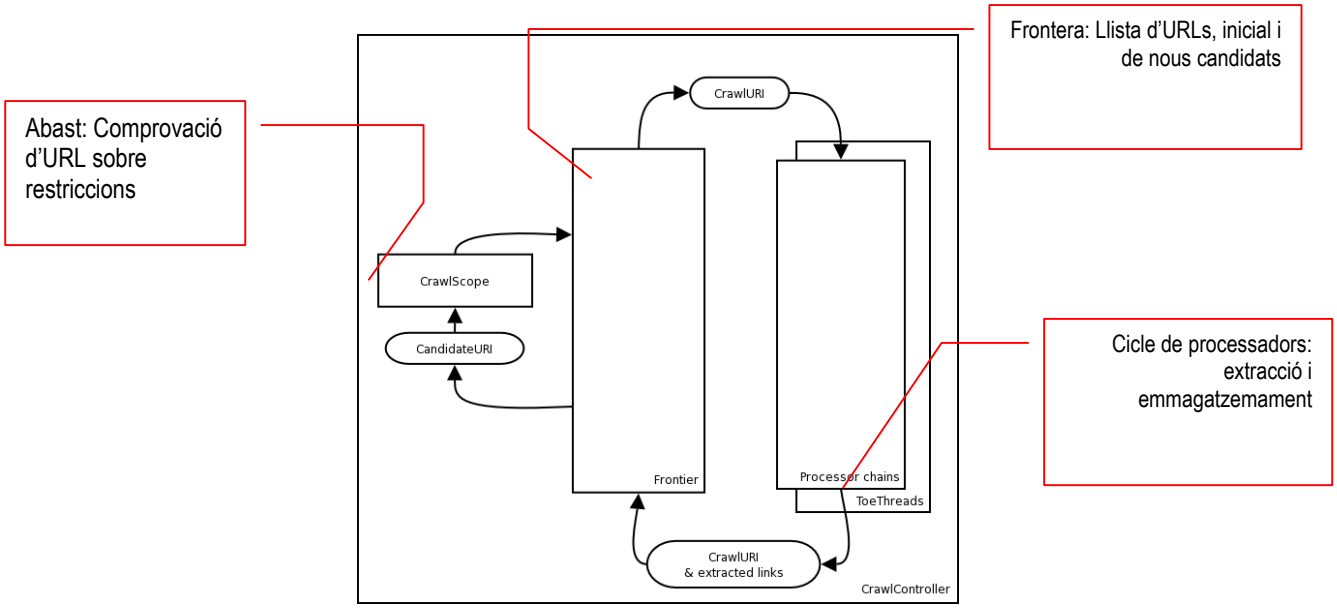
5.7. Testejar, avaluar i readaptar el programari fins assolir els objectius del projecte PADICAT.

Pel que fa al mòdul de captura d'Heritrix, el programari utilitza un llistat d'URL com a dada bàsica d'entrada. Durant la captura existeixen diversos mòduls interns que processen cada ítem del llistat, segons l'exemple que es mostra a continuació:

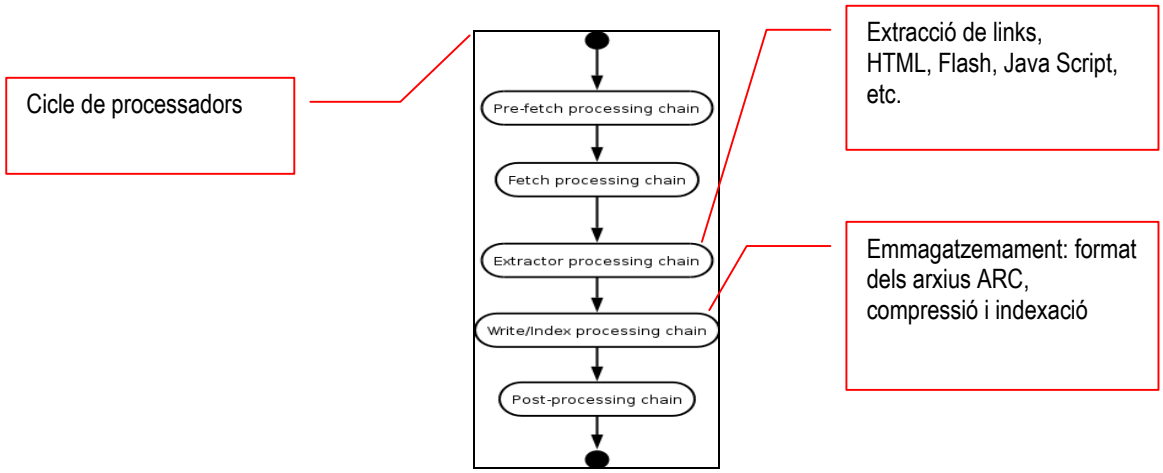
seed.txt

```
http://www.gencat.net/
http://web2.caib.es/owa/g0.frame_page2?codi=7
http://formacioprofessional.caib.es/web/index.htm
http://www.cult.gva.es/ivece/ivaqe/default_ivaqe.htm
http://www.ensenyament.com/
http://www.cisec.org/
http://www.expodidactica.com/
http://www.forumcatalunya.com/
http://www.xtec.net/fp/
http://usuarios.lycos.es/jgarciad/diners/
http://www.pisunyer.org/
```

El flux de treball intern inclou l'anàlisi de la llista d'entrada en el mòdul *frontera* (frontier), per comprovar si la URL és dins de les restriccions de captura. Vegeu el gràfic següent:



El cicle de processadors extrauen les dades de cada recurs digital, com s'escenifica al gràfic següent.



El procés d'extracció de links realitza la tasca de descobrir nous recursos que es trobin dins les pàgines web capturades, i agregar-los a la llista d'URL candidats per revisar-ne les restriccions. L'extracció de la URL es realitza en forma completa, és a dir, tots els URL que localitza el sistema són integrats a la llista de possibles candidats. Aquests recursos seran examinats en el procés de frontera per superar o no les restriccions d'abast del sistema, abans de decidir-se si el recurs ha de ser capturat, o no.

L'emmagatzemament de dades es realitza emprant el format de compressió d'arxius ARC, que permet desar diversos recursos dins d'un mateix arxiu comprimit. Aquest format permet la indexació dels arxius emmagatzemats, per a la seva cerca i visualització correcta.

L'arquitectura d'Heritrix permet que cada mòdul pugui ser implementat d'acord a les necessitats particulars del projecte. Per exemple, el mòdul d'emmagatzemament d'ARC s'ha integrat durant el test en una base de dades pròpia, com també la modificació dels mòduls de verificació d'abast i de frontera.

Pel que fa la les restriccions que s'apliquen al sistema per filtrar els recursos no desitjats, s'integren al sistema per mitjà d'un arxiu de configuració XML. Aquest arxiu conté els paràmetres de configuració de tots els mòduls implicats en la captura, així com en l'ordre en que es processen els diversos mòduls.

Es presenta a continuació un exemple d'arxiu de configuració.

order.xml

```
<crawl-order xsi:noNamespaceSchemaLocation="heritrix_settings.xsd">
<meta>
        <name>AllSeeds</name>
<description>Captura General</description>
<operator>Admin</operator>
<organization>Biblioteca Nacional de Catalunya</organization>
<audience/>
<date>20060126091407</date>
</meta>
<controller>
……
<map name="write-processors">
        <newObject
name="Archiver" class="org.archive.crawler.writer.ARCWriterProcessor">
<boolean name="enabled">true</boolean>
<map name="filters"> </map>
<boolean name="compress">true</boolean>
<string name="prefix">IAH</string>
<string name="suffix">${HOSTNAME}</string>
<integer name="max-size-bytes">100000000</integer>
        <stringList name="path">
<string>/var/arcs</string>
</stringList>
<integer name="pool-max-active">5</integer>
<integer name="pool-max-wait">300000</integer>
<long name="total-bytes-to-write">0</long>
</newObject>
</map>
……
</controller>
</crawl-order>
```

Configuració del procés d'escriptura per utilitzar el mòdul ARC

Per a cada mòdul existeix una secció dins de l'arxiu d'XML que permet configurar els paràmetres de cadascun dels mòduls

Heritrix permet crear treballs de captures amb restriccions totalment personalitzades, així com l'ordre de processament i els mòduls, que poden ser configurats d'acord a les necessitats de qualsevol projecte. Les proves realitzades a l'efecte, per tant, permeten preveure l'adequació del sistema a les necessitats generals, puntuals o variables del PADICAT.

5.8. Determinar junt amb el cap del projecte PADICAT de la Biblioteca de Catalunya les possibilitats tecnològiques de recollida de recursos electrònics catalans en línia.

5.8.1. Dominis/recursos internet de l'entorn català sobre els quals es realitzarà la recollida de recursos electrònics.

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 informació. 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.

Pel que fa a l'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[138] emprada habitualment pels membres del Laboratori d'Internet del CINDOC-CSIC, que servirà per definir el que genèricament entenem com a web:

> *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

[138] 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.

- 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 del que coneixem tradicionalment com *parts components*[139], 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 treballs[140] 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.

Pel que fa a l'abast temàtic, i com han recollit alguns autors[141], Internet està dissenyada per trencar les barreres geogràfiques i fer la informació accessible universalment. 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[142], on una part significativa del recurs hauria de ser:

---

[139] 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.

[140] 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]

[141] 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]

[142] "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]

- 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[143], que inclou:
  - 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[144], 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[145] 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

[143] 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]
[144] *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]

A partir d'aquests exemples, la informació electrònica susceptible de formar part del PADICAT és el grup de documents que contenen informació relativa a Catalunya o d'interès majoritari de la gent catalana, aspecte que conceptualment queda ja recollit a l'article 7 de la Llei de biblioteques de 1981[146]:

> *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. 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[147], o
- Webs ubicades a servidors de Catalunya[148], o
- Webs sota dominis geogràfics (.ES, .COM, .NET, .ORG, etc[149].) en llengua catalana[150], o

---

[145] 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]

[146] "Llei de biblioteques de Catalunya, de 24 d'abril de 1981". *Diari Oficial de la Generalitat de Catalunya*, núm. 123 (29 abr 1981).

[147] 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/

[148] 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.

[149] 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.

- Webs que no compleixin els requisits anteriors, però relacionades temàticament amb Catalunya[151]

5.8.2. Determinar les eines i mètodes per a obtenir automàticament els dominis/recursos de l'entorn català susceptibles de ser incorporats en la recollida de recursos.

El test ha permès conèixer el funcionament del programari Heritrix per la recollida automàtica dels recursos.

A partir de proposta bàsica inicial de perfils:

- Recopilació .CAT
- Recopilació llengua catalana
- Recopilació per IP
- Recopilació pàgines d'interès temàtic

Es resol que pel funcionament del capturador (Heritrix) és necessari crear un llistat inicial d'URL susceptibles de ser capturades. El procés descrit al punt 2.5 permet, a partir d'aquest pas, buidar altres URL que també seran susceptibles de formar part de la col·lecció, en base a les limitacions per a cada perfil (tasca o *job*)

Així, es va crear un primer llistat partint del perfil més senzill: el quart (temàtica). Es va implementar un mòdul d'extracció d'URLs per seus web categoritzades com a "llengua catalana" dins del directori del robot Google.

Bàsicament: es va crear un mòdul capaç d'extraure les URL d'aquest font, i buidar-los en la base de dades creada a l'efecte, amb l'objectiu de tenir en un format normalitzat un llistat a partir del qual procedir amb la resta de proves. En el relat de resultats s'entrarà en detall, però avancem que partíem d'aprox. 50.000 URL en llengua catalana[152], segons el directori de Google.

---

[150] 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.

[151] A partir de directoris, com Dmoz, Google, Yahoo, etc.

[152] El procés aleatori de verificació aporta que aprox. un 85% d'aquestes URL són efectivament en llengua catalana, no necessàriament creades a Catalunya.

Es realitzaren a continuació extensions a la interfície JMX[153], amb l'objectiu de facilitar les tasques més comunes: crear noves captures, controlar l'estat d'Heritrix, realitzar controls sobre els recursos capturats.

En un estadi inicial es va valorar completar aquest llistat amb URLs procedents d'altres directoris temàtics, com Nosaltres.com, etc. Finalment el factor temps va ser determinant per no incloure més fonts de partida.

En paral·lel a la prova, es va crear una base de dades en codi obert, MySQL, per gestionar els llistats d'URLs i els corresponents perfils.

5.8.3. Determinar les limitacions del model automàtic en l'obtenció de recursos electrònics de l'entorn català.

S'observa que Heritrix, contràriament al que s'havia pensat en abstracte, només pot partir d'un llistat d'URL concret. No pot, per tant, ser llançat directament contra la WWW per capturar "tot" el que trobi amb les limitacions que marquen els perfils.

Cal aprofundir en conseqüència en les possibilitats de crear un llistat inicial d'URLs que permetin formar una base sòlida a partir de la qual compilar recursos i valorar noves URLs que el procés d'extracció de links inclourà al cicle de processadors.

Les principals limitacions addicionals procedeixen de dos fronts:

- Errors de memòria (*out of memory*). Un dels problemes més recurrents és l'esgotament de la memòria. A causa del gran volum de dades processades, després de diversos dies de captura continuada, s'observen errors relacionats amb problemes de recursos d'administració ocasionats per la plataforma Java. Aquest tipus de problema, molt habituals a l'entorn Java, foren evitats reiniciant el procés de captura en el punt on es localitza l'error. Aquests problemes són de gran dimensió, però poden ser controlats reduint el nombre inicial d'URLs a capturar.

- Robots.txt. Aquest és un arxiu que molts administradors de pàgines web col·loquen a les seves pàgines i servidors. L'arxiu conté instruccions per als programes de captura, per mitjà de les

[153] Heritrix té implementat d'una forma parcial una interfície de control remot que utilitza JMX (Java Management eXtensions), que permet la programació i automatització del control de captura. Per exemple, enviant un arxiu comprimit amb la llista d'URLs (*seed.txt*) i l'arxiu de configuració (*order.xml*), és possible crear i controlar de manera completa tot el procés de captura. Malgrat que la interfície de control JMX no està desenvolupada per complet, les funcions de control remot més importants sí estan implementades, i és possible extendre les funcionalitats d'acord amb les necessitats del projecte. La interfície permet crear una arquitectura distribuïda, el que facilita poder tenir diverses màquines amb Heritrix

quals es limita l'accés a determinats continguts (per exemple, calendaris, continguts dinàmics que bloqueigen els cercadors, etc.). El problema està en el fet que moltes de les implementacions de l'arxiu Robots.txt només contenen instruccions perquè el robot no entri en cap apartat de la pàgina web. Caldrà, doncs, establir una solució de respecte a les exclusions, però tenint en compte una implementació creiem-ne que abusiva d'aquesta pràctica[154].

5.8.4. Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari de captació de recursos.

Com s'ha descrit, la metodologia passa per establir un primer llistat d'URLs, a partir de diversos perfils (possiblement temàtics, però no només), i en un nombre no superior als 25.000 URLs procedir a la captura sistemàtica de les pàgines web, amb la supervisió continuada del funcionament del motor de captura, així com de les propostes que el processador extrau de cada nova pàgina web.

Un dels conflictes més evidents durant les captures és la càrrega dels servidors on s'hostatgen les pàgines web que volem capturar. Els servidors ofereixen un servei que no sempre està preparat per una descàrrega continuada. Heritrix té el que anomenem paràmetres d'amabilitat (*politeness*) que representen temps d'espera que emulen, o imiten, el comportament humà en la navegació dins d'una pàgina web, evitant així la descàrrega de "tot, al mateix moment", i el conseqüent potencial col·lapse. Aquests paràmetres són modificables per l'administrador segons les necessitats del projecte.

Per defecte, Heritrix espera un temps configurable entre descàrrega i descàrrega de recursos d'un mateix servidor. Per aquest raó és necessari agrupar les seus web que es vulguis descarregar per evitar que Heritrix perdi temps entre les descàrregues. Per exemple: si El capturador es descarrega una imatge de la web de la Generalitat (www.gencat.net), no podrà descarregar cap altra fins que hagi passat el temps d'amabilitat estipulat (per defecte, entre un i dos minuts). Si la captura conté diversos URLs al llista de partida (*seeds*), Heritrix pot continuar amb la descàrrega d'altres recursos mentre no finalitzi el temps d'espera d'amabilitat.

Les estratègies de captura són emprades per agrupar recursos web que tenen una o més característiques similars. L'objectiu és crear conjunts d'URLs (*seeds*) que comparteixin el mateix perfil de configuració d'Heritrix, perquè es pugui realitzar una descàrrega sense pèrdua de temps d'espera per amabilitat. Una estratègia de captura pot contenir un, més d'un, o la combinació de diversos perfils

---

que realitzin les captures necessàries i conservin les dades en dipòsits comuns, arribant així a un alt grau de rendiment en l'escalabilitat i la resposta del sistema.

[154] El projecte danès no respecta l'arxiu robot.txt. Sí ho fa, per exemple, l'Internet Archive. Heritrix està preparat per respectar o no la instrucció, segons vulguin els administradors.

de configuració. Les estratègies de captura també són emprades com restriccions d'enllaç: per exemple, si una nova pàgina web és proposada durant la captura, es processa la seva informació per poder definir a quin perfil pertany. La llengua o la localització geogràfica poden ser algunes de les característiques dels perfils.

Les estratègies desenvolupades amb graus de suficiència han estat per llengua, i per localització geogràfica. Atès el baix nombre de dominis .CAT (quatre pàgines web, en el moment de realització del test), no s'ha inclòs en el procés d'estratègies.

Pel que fa a l'estratègia de llengua catalana, s'ha optat per provar d'identificar la llengua catalana en les metadades del recurs. Si l'etiqueta corresponent no existeix o es detecta anomalia, es realitza una prova en el contingut html de la pàgina, en base a un grup de paraules que tenen alta freqüència d'ús en llengua catalana, alhora que no presenten paral·lels en altres llengües: "aixo", "amb", "perque".

Pel que fa a l'estratègia de localització geogràfica, s'ha optat per comparar els resultats de la captura amb la base de dades externa que manté l'organització RIPE NCC. Aquest organisme té un registre amb la informació associada de les direccions d'Internet ubicades a Europa. Malgrat que la ubicació geogràfica no està necessàriament relacionada amb el contingut, aquesta segueix essent un factor molt important dins de la classificació de les pàgines web.

Cal definir noves estratègies de captura en base a futures proves del programari, ja en fase d'explotació.

5.8.5. Determinar els volums de recollida òptims per als testeigs.

La mida, el volum d'una pàgina web és una de les característiques més importants perquè està lligat estretament al temps de captura, i tanmateix amb els recursos d'emmagatzemament del sistema.

En tot cas, tenint en compte les dimensions del projecte que ens ocupa, no és una variable que hagi de condicionar el sistema de funcionament. La descàrrega de 100 GB procedent dels 25.000 URLs seleccionats va ocupar les dues màquines durant 15 dies. Això representaria, *grosso modo*, una captura (no completa) estàtica d'entre el 10 i el 25% de les pàgines web de Catalunya.

A partir de la prova realitzada, es va crear una senzilla classificació segons la mida de la pàgina web: recursos petits (menys d'1Mb), mitjans (1-100Mb) i grans (a partir de 100Mb). Observem que un 68% de les pàgines web compilades són petites, 30% mitjanes i 2% grans. Si extrapolem aquest resultat al que pot significar la web catalana, podem entendre que el 98% de les pàgines web de Catalunya tenen menys de 100Mb, essent la majoria d'aquestes (el 68% del total de la mostra) menors d'1 Mb.

5.8.6. Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari d'indexació de recursos.

El format d'emmagatzematge que empra el sistema és ARC, i el mòdul del programari que se n'ocupa és el BAT (Bnf Arc Tools), desenvolupat per la Bibliothèque Nationale de France. No és res més que un conjunt d'instruccions que permeten administrar i modificar arxius comprimits del tipus ARC, DAT, i CDX. Aquest mòdul es complementa amb el NutchWax (*Nutch Web Archive eXtensions*), un motor de cerca de codi obert desenvolupat per la fundació Apache, que té dues funcions bàsiques: d'una banda, crear índexs de les dades emmagatzemades; d'altra banda, permet la cerca en les dades del dipòsit, com ho faria un motor de cerca d'Internet.

Malgrat que la majoria de dades descarregades són elements en format HTML, i imatges, existeixen també altres tipus de formats que poden presentar incompatibilitat amb els sistemes indexadors, que en definitiva es basen en el text. Així i malgrat que es preveuen millores, els arxius PDF han de ser examinats a banda del sistema per a la seva posterior indexació. Arxius d'imatge, vídeo o àudio, etc., queden exclosos de la indexació.

5.8.7. Valorar i determinar el nombre i metodologia a seguir per als testeigs del programari de cerca i visualització de la base de dades generada a partir de la recollida i indexació dels recursos electrònics d'àmbit català.

El mòdul previst per a la visualització dels recursos de la col·lecció que forma el dipòsit és el Wera[155] (*Web Archive Access*) va ser desenvolupat inicialment per la Nasjonalbiblioteket (Noruega), i actualment s'hi ha afegit l'Internet Archive i la resta de membres de la comunitat. De fet és una interfície de cerca i visualització dels resultats que serveix per navegar, amb el motor de cerca NutchWax, dins d'un dipòsit de seus web.

Una de les limitacions observades són els links que apareixen encastats dins de continguts multimèdia, com Flash, Java Script, Aplicacions, etc., són una problemàtica bastant habitual, ja detallada en l'anàlisi d'altres projectes similars. L'extracció de les URLs encastades no és un problema. El problema rau en el format de visualització de les pàgines web conservades, per part de l'usuari final. No n'és factible la modificació automàtica i per tant dirigeixen l'usuari al link original, i no, com caldria, a la pàgina web conservada.

[155] http://archive-access.sourceforge.net/projects/wera/

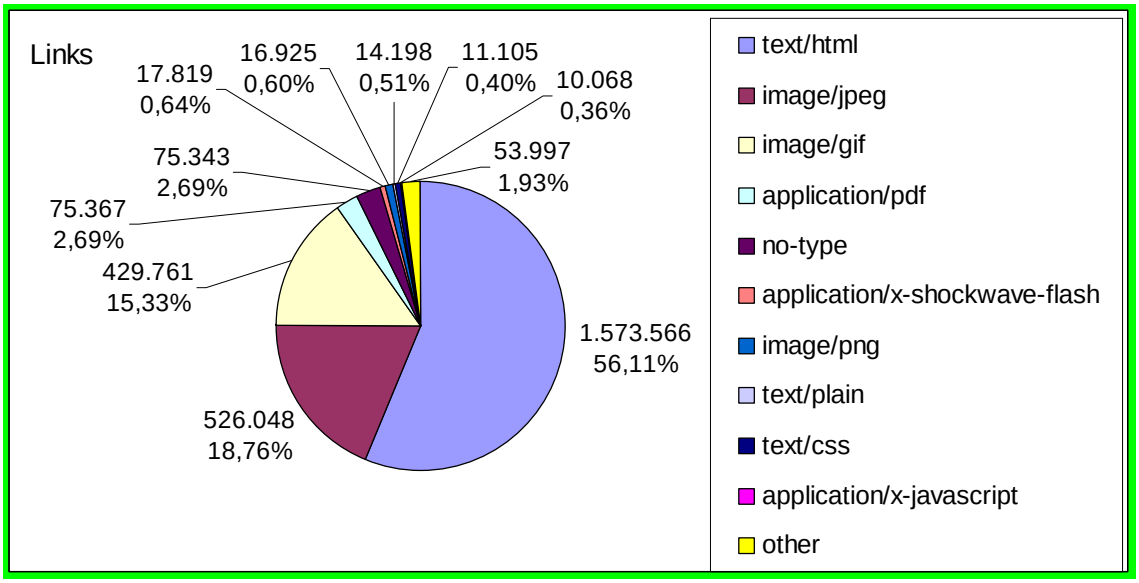
5.9. Recollir tota la informació resultant dels testeigs.

5.9.1. Objectius assolits i problemes detectats quant a l'obtenció automàtica de dominis/recursos en l'entorn català. Limitacions.

L'experiència més significativa, ja plantejada, fou a partir del conjunt de recursos "en llengua catalana" del directori Google. D'un total aproximat de 50.000 seus webs se'n van seleccionar la meitat, uns 25.000.

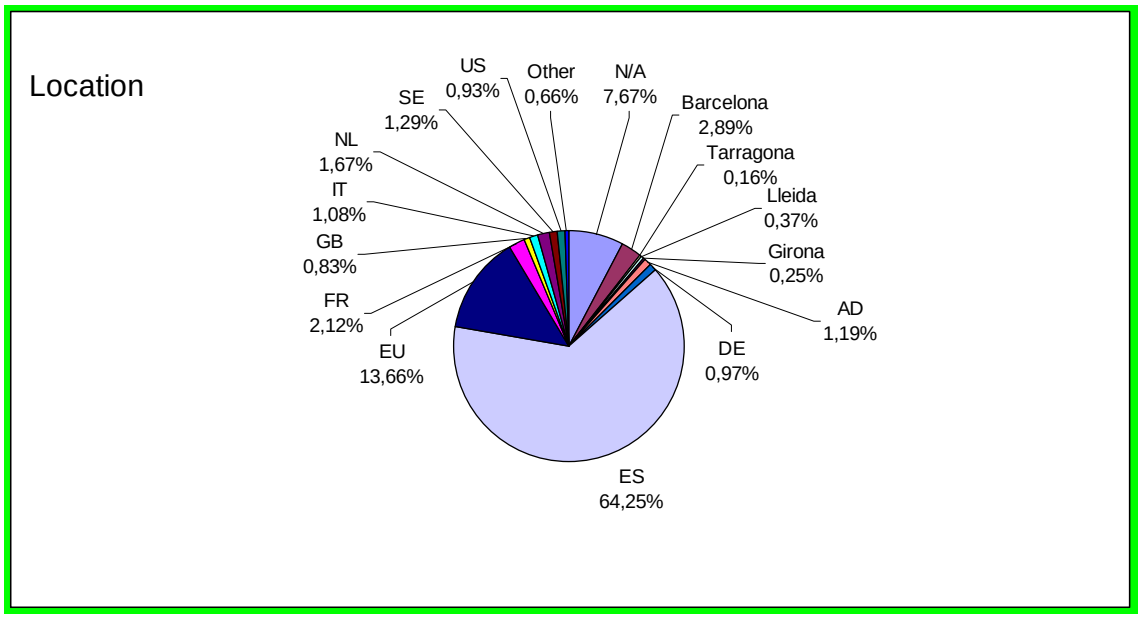
Amb aquest llistat d'URLs es va realitzar una prova de captura exhaustiva amb el capturador Heritrix. La captura total va trigar 15 dies, i el resultat permet assumir les dades de la descàrrega: uns 100 GB de dades, més de 2,5 milions d'arxius procedents dels 25.000 seus web.

Pel que fa a tipologia de formats, el 56% és text/html, les imatges estàndard representen un total aproximat del 36%, etc. Segons il·lustra el gràfic següent:



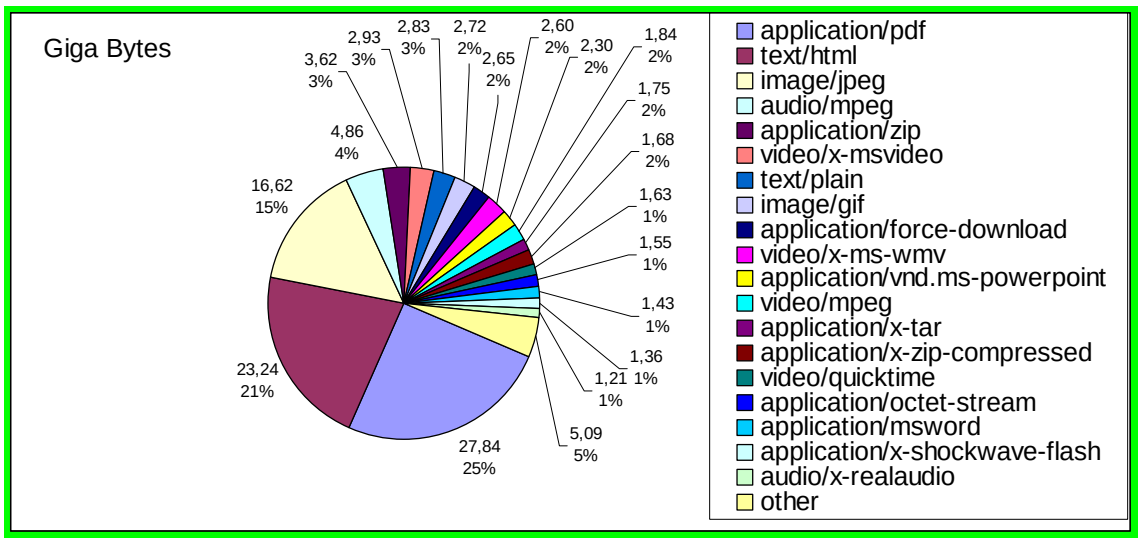
Les limitacions han estat relatades en l'apartat 2.6.3. del present informe.

Pel que fa a la ubicació geogràfica dels servidors, per control de les IP, els resultats no són determinants.

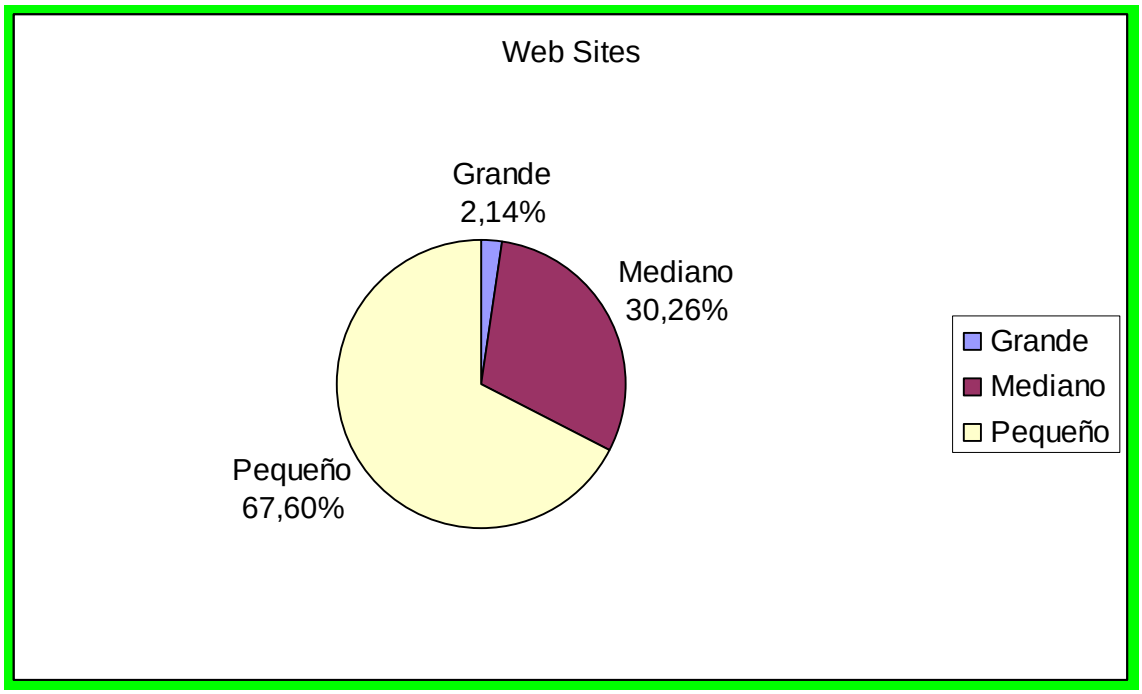


5.9.2. Volum de dades generat per a cada recollida.

Pel que fa a la mida segons el format d'arxiu, el tipus més voluminós és el pdf (25% del volum total, 27,84 GB), seguit de l'html (21%, 23,24 GB) i la imatge jpeg (15%, 16,62GB), segons s'entreveu del gràfic següent:



El tipus de pàgina web, com ja hem avançat, ens permet establir el tipus de recursos que conté la web catalana, tot fent prospecció a partir de la prova test. Sabem així que el 2,14% de les pàgines tenen una mida gran (més de 100Mb), el 30,26% una mida mitjana (1 a 100Mb) i el 67,60% de mida petita (menys d'1 Mb).



5.9.3. Temps dedicat a cada recollida segons el nombre de urls recollit i segons els maquinari i les limitacions utilitzades en el testeig.

Sense informació documentada.

5.9.4. Anàlisi valoratiu dels testeigs del programari de recollida, del programari d'indexació i del programari de cerca i visualització de la base de dades de recursos electrònics d'àmbit català: costos en temps i maquinari, càlcul de requeriments mínims segons els objectius. Problemes i limitacions observades.

Sense informació documentada.

5.9.5. Adaptacions de programari realitzades.

Sense informació documentada.

5.9.6. Requeriments necessaris a nivell de telecomunicacions.

Sense informació documentada.

5.10. Documentar tots els passos realitzats amb indicació de...

5.10.1. Programari instal·lat: versions, pedaços, errors, limitacions.

A partir de les conclusions relatades al punt 2.3. d'aquest informe es descarta invertir temps del test en una plataforma que no sigui la basada en Heritrix, i en conseqüència les proves es realitzen en base als quatre mòduls ja especificats en les versions disponibles a data del present treball:

- Heritrix (versió 1.6, desembre 2005)
- BAT (versió 0.07, desembre 2005)
- NucthWax (versió 0.4.0, octubre 2005)
- Wera (versió 0.4.0, octubre 2005)

5.10.2. Parametrització realitzada per a cada testeig.

Sense informació documentada.

5.10.3. Adaptacions realitzades per a cada testeig.

Sense informació documentada.

5.10.4. Objectius a assolir per a cada testeig.

Sense informació documentada.

5.10.5. Resultats obtinguts per a cada testeig.

Sense informació documentada.

5.10.6. Altres que es creguin necessaris en el moment de realitzar les proves.

El test s'ha realitzat a les dependències de la BC[156], amb dos PCs d'igual característiques:

- Processador Intel Pentium IV 3.2GH
- 2GB de memòria RAM
- Disc dur de 1,2 TB (3x400GB)
- Estació de treball de l'analista.

[156] Es preveu la utilització dels següents recursos de maquinari en la fase de producció.: 1 servidor[156] (SUN Fire V490 o similar), 1 GB memoria; 1 servidor[156] (SUN Fire V490 o similar), 1 GB memoria; 1 dipòsit[156] de disc de 10 TB (creixement anual previst 10 TB); 1 llibreria[156] de cintes LTO de 10 TB (creixement anual previst 10TB); 1 servidor estàndard per plataforma web.

5.11. Fer ús de les eines de consulta i contacte necessaris per a mantenir-se informat/da de les novetats de caràcter tecnològic que puguin sorgir en el període comprés en el contracte.

La participació en la comunitat Heritrix ha estat promoguda des de l'acció de consulta i contacte per mitjà de la llista de distribució <archive-webcrawler>[157]. Creada l'11 de febrer de 2003, hi participen membres de la majoria de biblioteques nacionals implicades en projectes similars, que de fet formen part de l'International Internet Preservation Consortium (IIPC), així com molt activament els creadors i desenvolupadors del programari, de l'organització Internet Archive, com Gordon Mohr o Michael Stack.

[157] http://sourceforge.net/projects/archive-crawler

6. Planificació a curt i mig termini i fases d'execució

6.1. Calendari gràfic

Fase Disseny: PRELIMINARS PADICAT (2005)	2005 jun	2005 jul-ago	2005 set	2005 oct	2005 nov	2005 des	2006 gen	2006 feb	2006 mar	2006 abr	2006 mai	2006 jun	2006 jul-ago	2006 set	2006 oct	2006 nov	2006 des
2005_05 Planificació fases d'execució																	
2005_01 Anàlisi de la qüestió mundial Models existents Conclusions a l'anàlisi																	
2005_02 Anàlisi i context Catalunya: recursos, agents implicats, aspectes legals																	
2005_06 Perfils i tasques recursos humans Perfils i tasques Calendari d'incorporació																	
2005_01b Intercanvi d'experiències Visita i entrevista BN Suècia [et al.] Assistència IWAW Viena (sep) Assistència *Jornadas* Madrid (oct)																	
2005_07 Estudi de costos fase de Producció																	
2005_03 Disseny del sistema d'informació Abast, circuit captura, organització i accés als recursos																	
2005_04 Anàlisi maquinari i programari i Test Maquinari ubicació, plataformes i opcions Programari existent Test Captura exhaustiva Test Captura per FTP Test Captura d'activitats concretes: *Nadal* Test Catalogació Test Preservació i gestió múltiple còpia																	

Fase Producció: PLA PILOT PADICAT (2006)	2006 gen	2006 feb	2006 mar	2006 abr	2006 mai	2006 jun	2006 jul-ago	2006 set	2006 oct	2006 nov	2006 des
2006_01 Captura sistemàtica											
2006_02 Processament tècnic: preservació Proves sistema doble còpia Proves emulacions											
2006_03 Processament tècnic: catalogació Metadades Automatització metadades											
2006_04 Comunicació Especialista Generalista											
2006_05 Captura d'esdeveniments *(Nadal, sant Jordi i els llibres, Ramadam a Catalunya, etc*											
2006_06 Gestió dels acords i Captures per acords, fase A CBUC+RACO Generalitat de Catalunya i entitats associades Ajuntaments Universitats Associacions Mitjans de comunicació											
2006_07 Web PADICAT estrena 11 setembre											
2006_08 Revisions del model: abast i organització projecte Model integral: replantejament criteris Model selectiu: replantejament criteris Model híbrid: calendari											
2006_09 Avaluació sistema i prospecció Captura Processament Preservació Accés											
2006_10 Disseny Oficina PADICAT											

Fase Explotació: OFICINA PADICAT (2007-2008)	2007 gen	2007 feb	2007 mar	2007 abr	2007 mai	2007 jun	2007 jul-ago	2007 set	2007 oct	2007 nov	2007 des	2008 gen	2008 feb	2008 mar	2008 abr	2008 mai	2008 jun-
2007_01 Entrada reforç catalogadors Metadades Automatització metadades Control accessibilitat																	
2007_02 Gestió dels acords i Captures per acords, fase B i C Generalitat de Catalunya i entitats associades Ajuntaments Universitats Associacions Mitjans de comunicació																	
2007_03 Captura sistemàtica																	
2007_04 Processament tècnic: preservació Proves sistema doble còpia Proves emulacions																	
2007_05 Procés retrospectiu Recerca Atenció al dipositant																	
2007_06 Captura esdeveniment: Eleccions 2007																	
2007_07 Rendiment òptim																	
2007_08 Avaluació i prospecció 2009-2011																	

6.2. Descripció dels processos

Fase Disseny: Preliminars Padicat (2005)

2005_05	Planificació fases d'execució
Calendari	01/06/2005 a 08/08/2005
Persones implicades	Cap de projecte Padicat (CP), Coordinadora General (CG)
Tasques principals	En base a la proposta adjudicada (expedient 07/05), planificació a curt i mig termini i fases d'execució
Resultat	Informe amb calendari i descripció dels processos 2005-juny 2008

2005_01	Anàlisi de la qüestió mundial
Calendari	01/06/2005 a 31/07/2005
Persones implicades	CP
Tasques principals	S'ha estudiat en profunditat els projectes de dipòsit nacional digital que existeixen arreu del món. Es detecten tendències i s'estableixen contactes amb responsables d'alguns dels projectes.
Resultat	Informe: capítol "Estat de la qüestió arreu del món"

2005_02	Anàlisi i context a Catalunya: recursos, agents implicats i aspectes legals
Calendari	01/07/2005 a 08/08/2005
Persones implicades	CP
Tasques principals	S'ha analitzat l'existència de projectes similars a Padicat. S'ha fet esment d'altres projectes resultat de la cooperació entre agències. Es presenta un llistat inicial d'agents a implicar en la fase pilot (2006). Es revisa la legislació existent a Espanya i a la resta de països on existeixen projectes de dipòsit nacional digital.
Resultat	Informe: capítol "Context a Catalunya"

2005_06	Perfils i tasques recursos humans
Calendari	01/07/2005 a 30/09/2005
Persones implicades	CP
Tasques principals	Es planificarà les necessitats del projecte en relació al personal que ha de formar part de l'equip de treball, amb anàlisi dels perfils i les tasques a realitzar, mètodes de contractació, i costos derivats de la contractació.
Resultat	Informe amb el detall dels perfils i tasques dels recursos humans de l'equip.

2005_01b	Intercanvi d'experiències
Calendari	01/09/2005 a 30/10/2005
Persones implicades	CP, CG
Tasques principals	En relació a l'epígraf 2005_01 es duran a terme contactes amb la resta de projectes existents. Es visitarà la seu d'algun dels projectes d'interès.
Resultat	Informe: capítol "Estat de la qüestió arreu del món", i tanmateix reports puntuals sobre el resultat o les previsions dels contactes i l'intercanvi de les experiències amb altres projectes.

2005_03	Disseny del sistema d'informació
Calendari	01/09/2005 a 31/12/2005
Persones implicades	CP, CG, TI (Tècnic informàtic)
Tasques principals	Es planifica el model a adoptar en el Padicat en relació a la definició de les fases del circuit: captura (abast i periodicitat del sistema), organització (gestió de metadades, d'acord amb els requisits mínims que la BC ha treballat en el CBUC), i accés als recursos.
Resultat	Informe amb disseny del sistema, requeriments tècnics i model de dipòsit a crear

2005_07	Estudi de costos fase de Producció
Calendari	01/09/2005 a 31/10/2005
Persones implicades	CP, CG

Tasques principals	Es planifica i es detalla les costos previstos curt i mig termini i principalment en el període 2006-2008 (fases d'execució).
Resultat	Informe amb estudi de costos al detall per al període 2006-2008

2005_04	Anàlisi maquinari i programari i Test
Calendari	01/10/2005 a 31/12/2005
Persones implicades	CP, TI, ATI (Personal Àrea de Tecnologia de la Informació)
Tasques principals	Identificació del maquina i programari òptims per a l'execució del projecte, tenint en compte les variables a tenir en compte en la selecció de les eines: mida de la web a capturar, periodicitat amb què es vol realitzar la captura, i nivell de fidelitat respecte a l'original (*look & feel*).
Resultat	Informe de l'anàlisi del maquinari i el programari emprat durant el test, així com les recomanacions de futur a adoptar pel projecte.

Fase Producció: Pla Pilot Padicat (2006)

2006_01	Captura sistemàtica
Calendari	01/01/2006 a 31/04/2006
Persones implicades	CP, TI, ATI (Personal Àrea de Tecnologia de la Informació)
Tasques principals	En base a l'informe que contempla el disseny del sistema, i tanmateix en relació al text del maquinari i programari, el pla pilot ha de tenir la capacitat de capturar automàticament els recursos digitals que s'incloguin en les fórmules utilitzades.
Resultat	Webs capturades i dades relatives al tipus de web.

2006_02	Processament tècnic: preservació
Calendari	01/02/2006 a 31/03/2006
Persones implicades	CP, IT
Tasques principals	La preservació dels llocs web inclou la conservació de l'estat (*look & feel)* de la web, així com les

	capacitats d'accedir als seus continguts tal com varen ser creats.
Resultat	Informe amb avaluació de la preservació, les pèrdues, els riscos i les accions a emprendre per garantir l'accés permanent.

2006_03	Processament tècnic: catalogació
Calendari	01/02/2006 a 31/03/2006
Persones implicades	CP, C1, C2 (Catalogadors)
Tasques principals	En base al disseny del sistema Padicat caldrà establir els circuits de catalogació per garantir la correcta i variada recuperació dels recursos que formin la col·lecció. Tanmateix, l'ús de les normes i recomanacions estàndards i els llenguatges que permetin un índex elevat d'autodescripció dels recursos, com són les metadades.
Resultat	Informe amb indicadors de rendiment en la catalogació, recuperació de la informació, i recomanació de millora en l'ús dels llenguatges descriptius.

2006_04	Comunicació
Calendari	01/06/2006 a 28/02/2006 i 01/09/2006 a 30/09/2006
Persones implicades	CP, CG, Comunicació BC
Tasques principals	Gestionar una correcta comunicació a la societat: gran públic, agents implicats, responsables polítics, per les vies generalistes i especialitzades, oferint una imatge dinàmica del recurs i la Biblioteca de Catalunya.
Resultat	Articles de divulgació, comunicats als mitjans de comunicació, entrevistes als responsables, presència en els debats sobre tecnologies i preservació.

2006_05	Captura d'esdeveniments
Calendari	01/03/2006 a 30/04/2006 i 01/10/2006 a 30/11/2006
Persones implicades	CP, CG, IT, C1
Tasques principals	Decisió , captura controlada i seguiment dels recursos digitals al voltant d'un esdeveniment concret, per determinar, de l'any 2006.

Resultat	Centre d'interès Padicat, i informe del rendiment del cas seleccionat

2006_06	Gestió dels acords i Captures per acords, fase A
Calendari	01/03/2006 a 31/12/2006
Persones implicades	CG, CP, GA (Gestor d'acords), IT, ATI
Tasques principals	Identificació dels agents, contacte, formalització de l'acord i seguiment del dipòsit voluntari d'una sèrie d'agents productor de la web catalana, susceptibles de dotar al fons del projecte d'una variada i qualitativament rica col·lecció de recursos.
Resultat	Canals de dipòsit voluntari, informe de rendiment.

2006_07	Web Padicat
Calendari	01/05/2006 a 31/12/2006
Persones implicades	CP, per definir responsabilitat webmaster
Tasques principals	En relació a la comunicació, es proposa prepara un lloc web propi del projecte Padicat, encabit en l'espai BC, que es presenti l'11 de setembre de 2006.
Resultat	Web, accés a les col·leccions que es puguin oferir, comunicació.

2006_08	Revisions del model: abast i organització projecte
Calendari	01/06/2006 a 31/08/2006
Persones implicades	CP
Tasques principals	En base al model previst en la fase de planificació del projecte, revisió dels paràmetres de valoració i proposta de reorientació del projecte.
Resultat	Informe amb anàlisi de les variables i orientacions de redefinició.

2006_09	Avaluació sistema i prospecció
Calendari	01/09/2006 a 30/11/2006

Persones implicades	CP, IT
Tasques principals	En base al funcionament ordinari del projecte, valoració global del pla pilot i estratègies d'augment del rendiment per a la creació de l'Oficina Padicat.
Resultat	Informe d'avaluació de pla pilot i propostes de millora.

2006_10	Disseny Oficina Padicat
Calendari	01/11/2006 a 31/12/2006
Persones implicades	CP, CG
Tasques principals	En relació a la resta de fases del pla pilot, disseny de l'oficina Padicat per als 2 anys d'explotació del projecte.¡, amb indicació de recursos personals i econòmics i viabilitat de la proposta existent.
Resultat	Informe amb disseny de l'Oficina Padicat 2007-2008.

Fase Explotació: Oficina Padicat (2007-2008)

2007_01	Entrada reforç catalogadors
Calendari	01/01/2007 a 28/02/2007
Persones implicades	CP, CG, C1, C2, C3
Tasques principals	Reforç de l'equip de catalogadors de l'Oficina Padicat i fixació d'objectius.
Resultat	Informe amb necessitats de personal i perfils i tasques, així com valoració d'emprendre estratègies paral·leles de col·laboració en l'Oficina (estades en pràctiques, externalització, etc.)

2007_02	Gestió dels acords i Captures per acords, Fase B i C
Calendari	01/01/2007 a 31/12/2008
Persones implicades	CG, CP, GA (Gestor d'acords), IT, ATI
Tasques principals	Identificació dels agents, contacte, formalització de l'acord i seguiment del dipòsit voluntari d'una sèrie d'agents productors de la web catalana, susceptibles de dotar al fons del projecte d'una variada i qualitativament rica col·lecció de recursos. Aquest segon i tercer bloc d'agents complementaran la primera fase, que ocupa el pla pilot.

Resultat	Canals de dipòsit voluntari, informe de rendiment.

2007_03	Captura sistemàtica
Calendari	01/02/2007 a 30/04/2007 i 01/02/2008 a 30/04/2008
Persones implicades	CP, TI, ATI (Personal Àrea de Tecnologia de la Informació)
Tasques principals	En base a l'informe que contempla el disseny del sistema, i tanmateix en relació al text del maquinari i programari, el pla pilot ha de tenir la capacitat de capturar automàticament els recursos digitals que s'incloguin en les fórmules utilitzades.
Resultat	Webs capturades i dades relatives al tipus de web.

2007_04	Processament tècnic: preservació
Calendari	01/03/2007 a 31/05/2007 i 01/03/2008 a 31/05/2008
Persones implicades	CP, IT
Tasques principals	La preservació dels llocs web inclou la conservació de l'estat (*look & feel)* de la web, així com les capacitats d'accedir als seus continguts tal com varen ser creats.
Resultat	Informe amb avaluació de la preservació, les pèrdues, els riscos i les accions a emprendre per garantir l'accés permanent.

2007_05	Procés retrospectiu
Calendari	01/04/2007 a 30/06/2007 i 01/04/2008 a 30/04/2008
Persones implicades	CP
Tasques principals	Adquisició de recursos anteriors al pla pilot del projecte que tinguin especial interès per a la col·lecció del Padicat., en còpies de seguretat, captures privades, etc.
Resultat	Webs dipositades i dades relatives al tipus de web

2007_06	Captura esdeveniments

Calendari	01/04/2007 a 31/05/2007 i data 2008 per determinar.
Persones implicades	CP, CG, IT, C1
Tasques principals	Decisió , captura controlada i seguiment dels recursos digitals al voltant d'un esdeveniment concret, per determinar, de l'any 2007 i 2008. Per a l'any 2007, seguint l'exemple d'altres projectes, es recomana fixar l'acció en les eleccions al Parlament de Catalunya.
Resultat	Centre d'interès Padicat, i informe del rendiment del cas seleccionat

2007_07	Rendiment òptim
Calendari	01/04/2007 a 31/05/2007
Persones implicades	CP, CG, IT
Tasques principals	Després d'una fase de planificació (2005) i de producció (2006), així com la primera etapa d'explotació (2007) a mans de l'Oficina Padicat, es mesurarà el rendiment òptim del projecte, així com la dotació de personal i de recursos tecnològics, a fi de preveure els resultats a mig i llarg termini.
Resultat	Informe de rendiment òptim i visió a 10 anys.

2007_08	Avaluació i prospecció 2009-2011
Calendari	01/05/2008 a 30/06/2008
Persones implicades	CP, CG, IT
Tasques principals	En base al funcionament ordinari del projecte, valoració global del pla pilot i estratègies d'augment del rendiment per a la següent etapa, de consolidació, de l'Oficina Padicat.
Resultat	Informe d'avaluació de l'Oficina i propostes de millora.

7. Recursos humans necessaris per executar el projecte: perfils i tasques

7.1. Introducció

En la següent descripció es parteix de la base que per al projecte es compta amb l'expertesa i suport del personal de la BC, així com dels possibles socis de projecte.

És previsible que el nivell de dedicació dels diversos perfils proposats variï significativament al llarg del període 2006-2008.

Es presenta per tant unes necessitats de personal[158] a màxims, especialment pel que fa als de perfil tecnològic. Possiblement una o dues persones de perfil analista polivalent, en base al que es presentarà, sigui adient al projecte, per la impossibilitat actual de determinar l'ocupació real de cadascun dels perfils següents:

- Cap de projecte
- Analista especialitzat, dedicat preferentment a la recollida de recursos
- Analista especialitzat, dedicat preferentment a la gestió-organització dels recursos
- Dissenyador d'interfície web
- Bibliotecari de suport I
- Bibliotecari de suport II
- Bibliotecari especialista metadades I
- Bibliotecari especialista metadades II
- Administratiu

[158] Algunes de les fonts consultades són: *eCatàleg, portal de compres de l'Administració pública* (http://www.ecataleg.cat365.net/), *Departament d'Universitats, Recerca i Societat de la Informació* (http://www10.gencat.net/dursi) i el *VI Conveni col·lectiu únic d'àmbit de Catalunya del personal laboral de la Generalitat de Catalunya per al període 2004-2008* (http://www.gencat.net/governacio-ap/administracio/pdf/VIconveni.pdf)

Cap de projecte	
Perfil	
Lloc de treball	Tècnic superior A24
Acadèmic	Llicenciat en Documentació
Formació addicional	Màster o postgrau en gestió de recursos digitals
Professional	Experiència en projectes similars
Principals competències genèriques	Orientació a l'usuari Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Iniciativa Flexibilitat Sentit analític i crític
Principals competències professionals	Comunicació Anglès Direcció i gestió Medi professional I+D Gestió de la col·lecció Legislació Administració
Tasques	
Fase de Disseny (2005)	Planificació fases d'execució Perfils i tasques recursos humans Anàlisi estat de la qüestió mundial Anàlisi context a Catalunya: recursos, agents, aspectes legals Disseny sistema d'informació Anàlisi Maquinari i programari Test maquinari i programari Relacions agents internacionals Definició model: abast, organització, presentació
Fase de Producció (2006)	Coordinació equip tècnic Aliances PADICAT Difusió en entorns professionals Relacions agents internacionals Avaluació disseny sistema Avaluació web
Fase d'Explotació (2007-2008)	Coordinació equip tècnic Coordinació posada en funcionament sistema PADICAT Aliances PADICAT Difusió genèrica del projecte Difusió en entorns professionals Relacions agents internacionals Impuls del rendiment òptim Plantejament de millores Propostes de preservació digital Procés retrospectiu Avaluació i prospecció

Analista especialitzat, dedicat preferentment a la recollida de recursos	
Perfil	
Lloc de treball	Analista informàtic A1
Acadèmic	Enginyeria o llicenciatura
Professional	Experiència en projectes similars
Principals competències genèriques	Orientació a l'usuari Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Flexibilitat Sentit analític i crític Documentació de processos
Principals competències professionals	Estàndard W3C Tecnologies i protocols web Experiència en l'anàlisi i millora del programari en codi obert Experiència en la documentació de projectes Anglès tècnic Experiència en configuració de maquinari amb sistema operatiu Linux Experiència en configuració de paquets de programari sobre sistema operatiu Linux Experiència en programació i adaptació de programari en Java
Tasques	
Fase de Producció (2006)	Adaptació del programari en la recollida de recursos web Parametrització i adaptacions del programari Documentació dels processos Mantenir-se informat novetats per mitjà eines de consulta i contacte amb al comunitat Millora rendiment maquinari i programari
Fase d'Explotació (2007-2008)	Adaptació del programari en la recollida de recursos web Parametrització i adaptacions del programari Documentació dels processos Mantenir-se informat novetats per mitjà eines de consulta i contacte amb al comunitat Optimitzar rendiment maquinari i programari Avaluació i prospecció

Analista especialitzat, dedicat preferentment a gestió-organització dels recursos	
Perfil	
Lloc de treball	Analista informàtic A1
Acadèmic	Enginyeria o llicenciatura
Professional	Experiència en projectes similars
Principals competències genèriques	Orientació a l'usuari Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Flexibilitat Sentit analític i crític Documentació de processos
Principals competències professionals	Estàndard W3C Tecnologies i protocols web Experiència en l'anàlisi i millora del programari en codi obert Experiència en la documentació de projectes Anglès tècnic Experiència en configuració de maquinari amb sistema operatiu Linux Experiència en configuració de paquets de programari sobre sistema operatiu Linux Experiència en programació i adaptació de programari en Java
Tasques	
Fase de Producció (2006)	Adaptació del programari en la gestió dels recursos web recopilats Documentació dels processos Mantenir-se informat novetats per mitjà eines de consulta i contacte amb al comunitat Millora rendiment maquinari i programari
Fase d'Explotació (2007-2008)	Adaptació del programari en la gestió de recursos web recopilats Documentació dels processos Mantenir-se informat novetats per mitjà eines de consulta i contacte amb al comunitat Optimitzar rendiment maquinari i programari Formular estratègies de preservació digital Avaluació i prospecció

Dissenyador d'interfície web	
Perfil	
Lloc de treball	Programador informàtic B
Acadèmic	Diplomatura o enginyeria tècnica
Professional	Experiència en projectes similars
Principals competències genèriques	Orientació a l'usuari Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Flexibilitat Sentit analític i crític Documentació de processos
Principals competències professionals	Disseny web Usabilitat Tecnologies i protocols web Experiència en l'anàlisi i millora del programari en codi obert Experiència en la documentació de projectes Anglès tècnic Experiència en configuració de maquinari amb sistema operatiu Linux Experiència en configuració de paquets de programari sobre sistema operatiu Linux Experiència en programació i adaptació de programari en Java
Tasques	
Fase de Producció (2006)	Proposta d'arquitectura i interfície web per al dipòsit voluntari de web Proposta de disseny web corporativa Usabilitat del sistema Documentació dels processos Producció web corporativa
Fase d'Explotació (2007-2008)	Producció del sistema de dipòsit voluntari via web Optimitzar rendiment maquinari i programari Avaluació i prospecció

Bibliotecari de suport I	
Perfil	
Lloc de treball	Bibliotecari B18
Acadèmic	Diplomatura en Biblioteconomia i Documentació
Professional	Experiència en gestió de col·leccions
Principals competències genèriques	Treball en equip Orientació als resultats Planificació i organització Sentit analític i crític
Principals competències professionals	Gestió de la col·lecció Gestió de continguts Anàlisi i representació documental Descripció i organització documental Recuperació de la informació
Tasques	
Fase de Producció (2006)	Suport al cap de projecte Seguiment captures automàtiques Seguiment captures per acords Documentació de processos i resultats
Fase d'Explotació (2007-2008)	Suport al cap de projecte Seguiment captures automàtiques Seguiment captures per acords Seguiment captures d'esdeveniments Preservació digital Avaluació i prospecció

Bibliotecari de suport II	
Perfil	
Lloc de treball	Bibliotecari B18
Acadèmic	Diplomatura en Biblioteconomia i Documentació
Professional	Experiència en gestió de col·leccions
Principals competències genèriques	Treball en equip Orientació als resultats Planificació i organització Sentit analític i crític
Principals competències professionals	Gestió de la col·lecció Gestió de continguts Anàlisi i representació documental Descripció i organització documental Recuperació de la informació
Tasques	
Fase de Producció (2006)	Suport al cap de projecte Seguiment captures automàtiques Seguiment captures per acords Documentació de processos i resultats
Fase d'Explotació (2007-2008)	Suport al cap de projecte Seguiment captures automàtiques Seguiment captures per acords Seguiment captures d'esdeveniments Preservació digital Avaluació i prospecció

Bibliotecari especialista metadades I	
Perfil	
Lloc de treball	Bibliotecari B18
Acadèmic	Diplomatura en Biblioteconomia i Documentació
Professional	Experiència en anàlisi documental
Principals competències genèriques	Orientació a l'usuari Treball en equip Orientació als resultats Sentit analític i crític
Principals competències professionals	Anàlisi i representació documental Descripció i organització documental Recuperació de la informació
Tasques	
Fase de Producció (2006)	Catalogació dels recursos per metadades Documentació de processos i resultats
Fase d'Explotació (2007-2008)	Catalogació dels recursos per metadades Documentació de processos i resultats Avaluació i prospecció

Bibliotecari especialista metadades II	
Perfil	
Lloc de treball	Bibliotecari B18
Acadèmic	Diplomatura en Biblioteconomia i Documentació
Professional	Experiència en anàlisi documental
Principals competències genèriques	Orientació a l'usuari Treball en equip Orientació als resultats Sentit analític i crític
Principals competències professionals	Anàlisi i representació documental Descripció i organització documental Recuperació de la informació
Tasques	
Fase de Producció (2006)	Catalogació dels recursos per metadades Documentació de processos i resultats
Fase d'Explotació (2007-2008)	Catalogació dels recursos per metadades Documentació de processos i resultats Avaluació i prospecció

Administratiu	
Perfil	
Lloc de treball	C13
Acadèmic	Batxillerat, FP2 o mòdul sup. o equivalent
Principals competències genèriques	Orientació a l'usuari Treball en equip Orientació als resultats Sentit analític i crític
Principals competències professionals	Redacció correspondència Redacció i tramitació convenis
Tasques	
Fase de Producció (2006)	Suport al cap de projecte Redacció i tramitació dels convenis de cooperació amb els editors web Gestió de la correspondència
Fase d'Explotació (2007-2008)	Suport al cap de projecte Redacció i tramitació dels convenis de cooperació amb els editors web Gestió de la correspondència

8. Estudi de costos vinculats a la fase de producció

8.1. El projecte PADICAT (Patrimoni Digital de Catalunya): els antecedents

Les tecnologies de la informació i la comunicació han facilitat que el patrimoni cultural, científic i la informació en general es presentin en format digital.

Actualment, i tal com ho exposen les *Directrius per a la preservació del patrimoni digital* (Unesco, 2003), la majoria dels 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'altre tipus, es generen directament digitalment, o es digitalitzen a partir de material analògic ja existent.

Els productes que d'antuvi es generen digitalment no existeixen en un altre format que no sigui l'electrònic original.

Aquesta realitat, sumada a la voluntat de les persones, les institucions i els governs de vetllar per la preservació de qualsevol forma de patrimoni, ha possibilitat que les administracions de diversos països hagin endegat polítiques destinades a garantir l'accés permanent a la producció digital —la seva recopilació i emmagatzematge, el tractament, la preservació i la difusió--.

A Catalunya el Pla de Serveis i Continguts de la Secretaria de Telecomunicacions i Societat de la Informació del DURSI preveu accions com la que es presenta i el projecte PADICAT, de la mà de la Biblioteca de Catalunya, serà un recurs de referència a les comunitats ciutadana, científica, i mediàtica.

8.2. Arxivant la web

Com s'ha apuntat, i malgrat les dificultats, diversos països han entès la necessitat de passar a l'acció, i d'establir polítiques i emprendre accions de preservació per assegurar la pervivència de la producció digital, com ja s'havia fet històricament amb els documents impresos i en suports tradicionals, mitjançant les lleis nacionals del dipòsit legal. En la majoria dels casos que ja existeixen, ha estat la biblioteca nacional, amb diversos socis tecnològics, qui ha liderat el procés d'accés permanent al patrimoni digital.

El ventall de projectes existents contempla de manera desigual què cal prioritzar, i amb quines condicions: pàgines web, diaris digitals, weblogs, votacions electròniques, etc.

Però les dificultats són notables. Per començar, els mètodes tradicionals de preservació de la producció bibliogràfica (com el dipòsit legal) són de difícil aplicació en l'entorn digital perquè, a banda de la

possible obsolescència del text legal (el cas espanyol), els recursos digitals poden instal·lar-se en servidors d'arreu del món (com passa també amb els impressors que no són editors). Aquest fet dificulta la tradicional correspondència geogràfica entre la ubicació del productor i la llengua o la temàtica publicada. En segon lloc, la producció digital té un creixement exponencial, essent a més molt variable la durabilitat dels materials publicats a Internet (una pàgina web té una mitjana de 44 dies de vida) i, en conseqüència, limitada la possibilitat d'accés permanent. Finalment, la qüestió de la propietat intel·lectual del producte digital, que està mancada d'un dret basat en el principi de còpia per a la preservació que asseguri la conservació i perdurabilitat del patrimoni digital, amb les limitacions comercials que siguin necessàries.

8.3. Casos existents

Les experiències que existeixen, unes 20 a tot el mon, es poden classificar en tres models, segons l'abast de compilació dels recursos web: integral, selectiu, o híbrid.

Per un costat, el model integral o exhaustiu, característic de Suècia (1996), Noruega (2001), Finlàndia (1997) i Àustria (1999), que aposta per la integració automàtica del total de la web a partir de determinats criteris infraestructurals (lingüístics, segons el domini de les web, segons la ubicació del servidor, etc.).

Per un altre costat i assimilat per Austràlia (1996), Canadà (1994), Japó (2002) i el Regne Unit (2004), entre d'altres països, el model selectiu dirigeix les accions de recopilació d'acord amb una política selectiva temàtica (sobre un espai geogràfic determinat, al voltant d'un tema d'interès nacional, etc.); per fer-ho arriba a acords amb els editors o productors de recursos web.

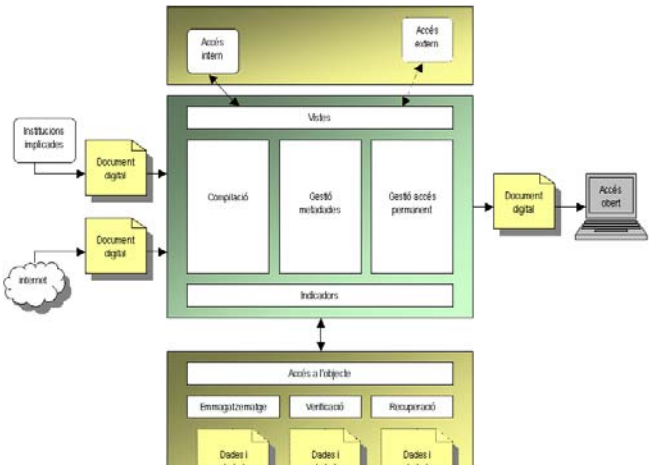
Aquests dos models han deixat pas en alguns països, però cada vegada amb més força, a models híbrids, com els de Dinamarca (1998), Nova Zelanda (1999) i França (2000), que complementen la captura periòdica del total de la web nacional, els acords amb els productors, i puntualment amb la compilació d'esdeveniments d'actualitat (eleccions, jocs olímpics, catàstrofes, etc.).

Addicionalment, existeix des de 1996 el gegant Internet Archive (http://www.archive.org), que compila sistemàticament *tota* la web mundial.

8.4. Arquitectura del sistema

El sistema no és tecnològicament complex, però sí requereix maquinari i capacitat de dipòsit.

Amb la fita estratègica de l'accés permanent al Patrimoni Digital de Catalunya, el sistema ha de contemplar les variables relacionades amb el cicle documental clàssic de les biblioteques i els serveis d'informació (adquisició + tractament + difusió).



Basat en el model Dulabahn (2004)

Una vegada definida la política de captura i els agents implicats que necessàriament caldrà implicar en el projecte, s'analitzarà amb deteniment les qüestions relacionades amb el funcionament intern del sistema, des de l'acció de captura fins a l'accés obert al document digital.

8.5. Calendari PADICAT 2005-2009

La Biblioteca de Catalunya ha iniciat el juny de 2005 les actuacions necessàries per impulsar amb garantia d'èxit el projecte PADICAT.

A partir de la recerca sobre els projectes existents, que ha inclòs entrevistes als responsables de sistemes anàlegs de Suècia, Dinamarca, i França, s'ha determinat optar pel model ja presentat com a híbrid, basat en la compilació dels recursos web catalans procedents de tres accions:

- Captura automàtica regular de les webs catalanes en base a determinats criteris infraestructurals (llengua, ubicació del servidor, domini *cat*, anàlisi d'enllaços web, etc.)
- Dipòsit voluntari sistematitzat, en base a acords amb institucions editores web, de tota índole (societats culturals, associacions de veïns, ONGs, ajuntaments, partits polítics, etc.)
- Compilació focalitzada i exhaustiva de determinats esdeveniments d'interès cultural, polític, social, o científic (eleccions al Parlament, descobriments científics, etc.)

El calendari d'execució, per al període 2005-2008, és el següent:

- 2005 Planificació
    - Anàlisi dels projectes i recursos existents, agents implicats i aspectes legals.
    - Definició de l'abast i test de maquinari i programari.
- 2006 Producció: Pla pilot
    - Captura global i sistemàtica (2 cops l'any) de la web catalana.
    - Acords per dipòsit voluntari regular amb 100 institucions.
    - Captura d'un esdeveniment d'interès nacional.
- 2007-2008 Explotació: Oficina PADICAT
    - Màxim rendiment del sistema.
- PADICAT, a 1 de gener de 2009
    - Eina informativa de referència a la societat catalana.
    - Projecte tecnològicament pioner a Espanya i de referència a Europa.
    - Implicació de 300 institucions de tot tipus en base a acords, segell PADICAT.
    - 100.000 webs, 50 milions d'arxius, 30 Terabytes de volum.
    - Accés en línia, en obert, a bona part de la col·lecció.

8.6. Plurianual PADICAT

En previsió de formalitzar l'acord amb el CESCA, es procedeix a especificar les dades relatives a les necessitats pressupostàries del projecte PADICAT (Patrimoni Digital de Catalunya), incloent la part repercutida que la BC aportarà al CESCA per ús de la seva infraestructura.

A la memòria del projecte (informe de costos), el cost global del PADICAT és sensiblement superior al que aquí s'exposa. I és així perquè el CESCA cedeix en bona mesura l'ús del seu equipament de supercomputació i emmagatzematge, i per tant l'aportació que assumeix la BC es redueix, especialment en la partida d'infraestructura.

8.6.1. Pressupost

Inversió de la BC	2006	2007	2008	Total per partides
Aportació al CESCA	66.000,00 €	70.000,00 €	75.000,00 €	211.000,00 €
Recursos humans BC	131.000,00 €	161.000,00 €	243.000,00 €	535.000,00 €
Promoció BC	2.000,00 €	8.000,00 €	10.000,00 €	20.000,00 €
Total per anys	199.000,00 €	239.000,00 €	328.000,00 €	766.000,00 €

8.6.2. Partida: aportació al CESCA

Inversió de la BC	2006	2007	2008	Total per partides
Aportació al CESCA	66.000,00 €	70.000,00 €	75.000,00 €	211.000,00 €
Recursos humans BC	131.000,00 €	161.000,00 €	243.000,00 €	535.000,00 €
Promoció BC	2.000,00 €	8.000,00 €	10.000,00 €	20.000,00 €
Total per anys	199.000,00 €	239.000,00 €	328.000,00 €	766.000,00 €

Aportació al CESCA	2006	2007	2008	Total per àrees
Infraestructura	30.000,00 €	32.000,00 €	35.000,00 €	97.000,00 €
Recursos humans	36.000,00 €	38.000,00 €	40.000,00 €	114.000,00 €
Total per anys	66.000,00 €	70.000,00 €	75.000,00 €	211.000,00 €

La BC realitza aquesta aportació directa al CESCA en concepte d'ús d'una part concreta de la infraestructura del CESCA i tanmateix per la dedicació de personal necessària per al bon funcionament del projecte.

Pel que fa a la infraestructura, l'aportació de la BC al projecte representa aproximadament el 22% del cost total en infraestructura. La resta, en una proporció aproximada del 78% del cost total, és assumida pel CESCA per a l'ús de la seva pròpia infraestructura per al PADICAT. Concretament es preveu l'ús de 2 servidors (SUN Fire V490 o similar), robot amb capacitat per a 10 TB de volum anuals, així com les màquines necessàries per a la creació i manteniment d'un portal web PADICAT, amb la corresponent interfície de consulta.

Pel que fa als recursos humans, la inversió es relaciona amb la dedicació d'un analista informàtic sènior AI-A (2006-2008), a temps complet, que haurà de dissenyar i garantir el funcionament del sistema.

8.6.3. Partida: recursos humans BC

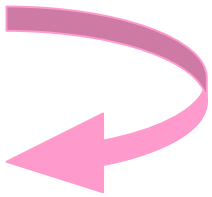
Inversió de la BC	2006	2007	2008	Total per partides
Aportació al CESCA	66.000,00 €	70.000,00 €	75.000,00 €	211.000,00 €
Recursos humans BC	131.000,00 €	161.000,00 €	243.000,00 €	535.000,00 €
Promoció BC	2.000,00 €	8.000,00 €	10.000,00 €	20.000,00 €
Total per anys	199.000,00 €	239.000,00 €	328.000,00 €	766.000,00 €

Recursos humans BC	2006	2007	2008	Total per àrees
Direcció	42.000,00 €	44.000,00 €	46.000,00 €	132.000,00 €
Tecnològics	- €	22.000,00 €	24.000,00 €	46.000,00 €
Bibliotecaris (2 a 4)	64.000,00 €	68.000,00 €	144.000,00 €	276.000,00 €
Administratiu	25.000,00 €	27.000,00 €	29.000,00 €	81.000,00 €
Total per anys	131.000,00 €	161.000,00 €	243.000,00 €	535.000,00 €

El capítol inclou la participació total de 6,5 persones, específicament per categories i per anys: 1 cap de projecte TS-A24 (2006-2008), ½ jornada programador informàtic PI-B (2007-2008), 1 bibliotecari catalogador B18(2006-2008), 1 bibliotecari catalogador B18 (2008), 1 bibliotecari de suport B18 (2006-2008), 1 bibliotecari de suport B18 (2008), i 1 administratiu C13 (2006-2008). A aquest personal cal sumar-hi conceptualment l'analista informàtic del CESCA.

8.6.4. Partida: promoció BC

Inversió de la BC	2006	2007	2008	Total per partides
Aportació al CESCA	66.000,00 €	70.000,00 €	75.000,00 €	211.000,00 €
Recursos humans BC	131.000,00 €	161.000,00 €	243.000,00 €	535.000,00 €
Promoció BC	2.000,00 €	8.000,00 €	10.000,00 €	20.000,00 €
Total per anys	199.000,00 €	239.000,00 €	328.000,00 €	766.000,00 €



Promoció BC	2006	2007	2008	Total per àrees
Disseny imatge logotip	1.000,00 €	- €	- €	1.000,00 €
Representació PADICAT	1.000,00 €	3.000,00 €	3.000,00 €	7.000,00 €
Producció materials	- €	5.000,00 €	7.000,00 €	12.000,00 €
Total per anys	2.000,00 €	8.000,00 €	10.000,00 €	20.000,00 €

El capítol inclou un partida mínima per a recursos que donin suport a la promoció i representació del servei, disseny del logotip (del projecte PADICAT, que s'oferirà a incloure en cada web que formi la col·lecció), publicació de banners o altres productes de promoció.