# CAT (CURATOR ARCHIVING TOOL): MILLORANT L'ACCÉS ALS ARXIUS WEB

**Ciro Llueca, Daniel Cócera**
Biblioteca de Catalunya (BC)
Barcelona, Espanya
padicat@bnc.cat

**Natalia Torres, Gerard Suades, Ricard de la Vega**
Centre de Supercomputació de Catalunya (CESCA)
Barcelona, Espanya
padicat@cesca.cat

**ABSTRACT**

PADICAT és l'arxiu web creat el 2005 a Catalunya (Espanya) amb l'objectiu de capturar, processar i donar accés permanent al patrimoni digital de Catalunya. Basa la seva estratègia de captura en el model híbrid (captura massiva del domini .cat; captura selectiva dels agents productors de les pàgines web catalanes; captura focalitzada d'esdeveniments públics). El sistema ofereix la seva col·lecció en obert, a Internet. Per fer-ho de manera òptima ha cregut necessari complementar els actuals programes de cerca i visualització amb una nova eina de programari lliure, CAT (Curator Archiving Tool), formada per tres mòduls orientats a gestionar eficaçment els processos de catalogació humana; publicar els recursos en directoris i centres d'interès temàtic; i oferir als usuaris informació estadística de valor afegit. En el marc de l'International Internet Preservation Consortium *meeting* (Viena 2010) es presenta a la comunitat internacional els avenços en la producció d'aquesta nova eina informàtica, i la filosofia que n'ha causat el disseny.

## 1. INTRODUCCIÓ

PADICAT (Patrimoni Digital de Catalunya) és l'arxiu web creat el 2005 a Catalunya (Espanya) per la Biblioteca Nacional de Catalunya (BC) i el Centre de Supercomputació de Catalunya (CESCA), amb l'objectiu de capturar, processar i donar accés permanent al patrimoni digital de Catalunya, entès com tota la producció cultural, científica i de caràcter general produïda en format digital[1] i publicada a Internet. La missió de PADICAT[2] és arxivar la Internet catalana.

La BC forma part de l'IIPC (International Internet Preservation Consortium), i com en la resta de projectes en funcionament, l'arxiu web amb seu a Barcelona es basa en l'aplicació d'una sèrie de programes informàtics que permeten la captura, l'emmagatzematge, l'organització, la preservació i l'accés permanent a una sèrie de versions de les pàgines web publicades a Internet per una comunitat determinada[3], en aquest cas la catalana, a Espanya.

A partir de l'anàlisi inicial dels models Internet Archive, Kulturarw3 (National Library of Sweden), Pandora (National Library of Australia), i Netarkivet.dk (The Royal Library, State & University Library, Denmark), i d'acord amb el que considerem que és la tendència generalitzada arreu de les biblioteques i arxius nacionals, el model de dipòsit que es persegueix a PADICAT és el model híbrid, consistent a:

- Compilar massivament els recursos digitals publicats en obert a Internet, per mitjà de la captura exhaustiva del domini .CAT.
- Impulsar el dipòsit sistemàtic de la producció web de les entitats catalanes, per mitjà de la identificació i el conveni amb les entitats i empreses representatives de la societat catalana.
- Promoure línies de recerca per mitjà de la integració temàtica dels recursos digitals de determinats esdeveniments de la vida pública catalana, com és el cas de les diverses campanyes electorals a Internet, el fenomen de la música en línia, o el món dels museus.

PADICAT utilitza en el seu funcionament ordinari Heritrix, Nutchwax, Wera, Wayback Machine, així com el programari

[1] Webb, C. *Guidelines for the Preservation of Digital Heritage*. United Nations Educational, Scientific and Cultural Organization, Paris, 2003. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

[2] See "What PADICAT is" at http://www.padicat.cat/en/quees.php

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

Web Curator Tool (WCT), desenvolupat per la National Library of New Zealand en col·laboració amb la British Library.

A setembre de 2010, PADICAT ofereix en el seu web (http://www.padicat.cat) accés en obert a 53.249 captures de 30.481 seus web i manté convenis de col·laboració amb 442 administracions públiques, empreses, universitats, associacions professionals i centres culturals de Catalunya. L'accés ofert en obert de tota la col·lecció, via cerca o navegació temàtica, és justament un dels punts forts de l'arxiu web català. Conjuntament amb l'objectiu de preservar la informació publicada a Internet, la visualització dels fons capturats a Internet ha estat una obsessió: conscients de les limitacions legals que afecten la majoria d'arxius web existents, PADICAT ha seguit des del seu començament la premissa de permetre l'accés obert a la seva col·lecció, defensada per l'Internet Archive, alhora que s'ha anat teixint artesanalment una xarxa d'entitats i empreses que autoritzen, per mitjà d'un conveni de col·laboració, la difusió de les versions capturades al portal de l'arxiu web.

L'objectiu estratègic de garantir una òptima recuperació i visualització de la informació capturada s'ha dut a terme per mitjà de l'assignació de valor afegit a les dades dels registres capturats, associant nova informació descriptiva als ítems digitals. En definitiva, catalogant els recursos web que PADICAT ha anat capturant o ha de capturar properament. A data d'avui s'han catalogat 17.700 pàgines web.

La informació proporcionada per la catalogació manual és la que haurà de permetre una millora substancial en la identificació i recuperació dels recursos segons el títol normalitzat, el posicionament alfabètic en el directori de recursos del repositori, la pertinença i la indexació dins d'unes categories temàtiques concretes, la interrelació dels propis recursos dins la col·lecció de PADICAT, etc. El repte últim és la intervenció en l'scoring dels actuals sistemes de cerca (Wera i Wayback) per a ponderar els resultats tenint en compte metadades procedents de la catalogació.

En definitiva, nous elements de cerca que han d'oferir com a resultat final una millor visibilitat dels recursos que formen part de la col·lecció, permetent a l'usuari final una interrogació lògica del sistema més enllà que conegui o no la URL del recurs desitjat. La catalogació de recursos web ha impulsat el disseny i producció d'un conjunt d'eines que anomenem CAT (Curator Archiving Tool).

És objectiu d'aquest informe presentat a l'International International Preservation Consortium *meeting* (celebrat en la 7th International Conference on Preservation of Digital Objects, iPRES2010) donar compte a la comunitat internacional de les millores plantejades per l'equip de PADICAT als processos de recuperació i visualització de la informació continguda en els arxius web, així com posar a disposició dels projectes homòlegs els avenços en aquestes línies de recerca i implementació.

## 2. CAT (CURATOR ARCHIVING TOOL)

És evident que la captura de pàgines web publicades a Internet està suficientment resolta pels programaris més madurs, com Heritrix. Per contra, els softwares d'indexació i els corresponents mòduls de visualització no han evolucionat al mateix ritme.

Com en la resta de catàlegs, la recuperació de la informació continguda en un arxiu web es realitza bé per mitjà de la cerca o bé per la navegació en un directori, habitualment temàtic. Només la part de cerca disposa de programaris específics força estesos a la comunitat, com són el Wera i el Wayback, a partir de consulta per mitjà d'una URL concreta, o d'una paraula clau. La navegació per directoris temàtics, plantejada en arxius web com Pandora o UK WAC, únicament pot sorgir quan existeixen processos tradicionals de catalogació, com el que se segueix també a PADICAT, amb l'objectiu últim d'integrar les seves col·leccions als catàlegs de la Biblioteca Nacional de Catalunya.

Així doncs, les etapes del procés de la gestió de recursos web com la captura o la cerca disposen de programaris prou madurs i molt usats per part de projectes similars, però existeix un buit en quant a programaris per a la catalogació i publicació d'aquests recursos .

Davant d'aquest repte, i per garantir una millor recuperació i visualització per part dels públics externs als arxius web, l'equip de PADICAT han dut a terme des de 2005 una sèrie d'assajos i implementacions que han fet possible, en l'actualitat, presentar una eina dirigida a influenciar positivament en els processos de cerca i recuperació d'informació, així com en la presentació d'aquesta informació al públic que consulta un arxiu web.

Es tracta dels mòduls que formen CAT (Curator Archiving Tool), dissenyats amb els criteris de modularitat, escalabilitat i internacionalització que garanteixen la seva publicació com a programari lliure.

S'ha dissenyat i s'aborda la producció d'un programari format per tres mòduls: el bàsic, de Catalogació (MOCA), que es complementa amb més funcionalitats a partir de mòduls addicionals: el de Publicació (MOPU), i el d'Estadístiques

(MOST), que units formen l'eina CAT (Curator Archiving Tool), objecte de la present comunicació.

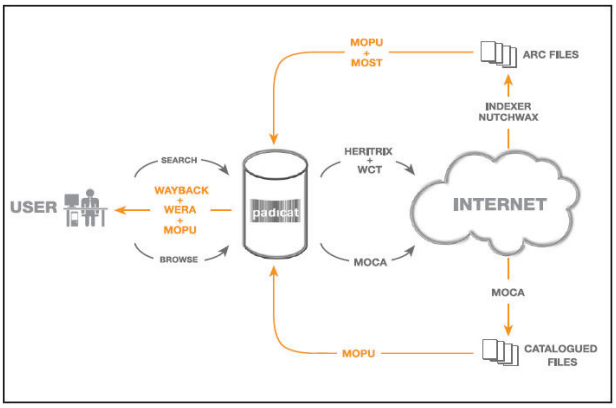


**Figura 1.** Programari utilitzat a PADICAT

Vegem a continuació una descripció dels mòduls que integren l'eina CAT (Curator Archiving Tool).

### 2.1. MOCA (Mòdul de Catalogació)

La catalogació de recursos es va iniciar utilitzant eines ja existents, essencialment el pioner Web Curator Tool (WCT), però amb l'experiència es van detectar mancances en el programari i necessitats particulars que van fer que ens plantegéssim desenvolupar un nou producte centrat en la catalogació.

Un dels principals requisits de MOCA és que permet l'enregistrament d'un històric de canvis en les metadades dels recursos catalogats ja que aquesta informació podria ser rellevant: l'exemple més evident és el canvi d'URL associada al recurs preservat.

La informació que es recull per catalogar els recursos es basa en el model de metadades Dublin Core, però es pot ampliar fàcilment sense alterar el model de dades. A PADICAT els recursos capturats són processats segons el sistema d'indexació[4] de la Biblioteca Nacional de Catalunya, en base a una classificació temàtica pròpia del portal PADICAT, i d'acord amb la procedència de la instrucció de captura (perquè pertany al domini .cat, als llocs webs procedents de convenis de col·laboració, a recursos recomanats, o a col·leccions monogràfiques).

[4] LEMAC, *Llistat d'Encapçalaments de Matèria en Català*, adaptació dels LCSH, *Library of Congress Subject Headings* amb aportacions procedents del Laval RVM, *Répertoire des Vedettes-Matière*, i el RAMEAU, *Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié*, que la BC manté.

Un control limita l'accés a l'eina als usuaris identificats per mitjà d'usuari i contrasenya, i una vegada validat l'usuari només té accés a les accions que pot realitzar segons el rol assignat. L'ús de rols simplifica les pàgines d'usuari i permet controlar acuradament les accions que pot realitzar cadascun d'ells. S'han definit tres rols basant-se en els perfils professionals del personal implicat en el procés de catalogació: el gestor (tasques de manteniment de l'eina), el catalogador (introducció i actualització dels recursos catalogats) i l'observador (control dels recursos catalogats).

El flux de treball de MOCA obliga a verificar la no existència d'un recurs ja integrat en el catàleg per evitar la duplicació de registres. Un cop superat el primer pas es requereix l'entrada de les metadades mínimes obligatòries i l'ordre d'emmagatzematge.

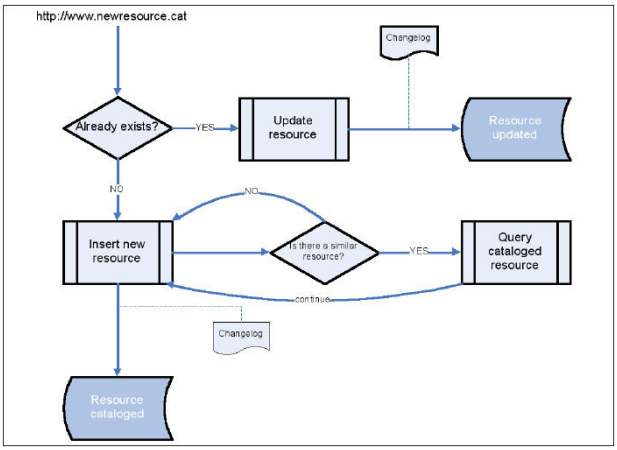


**Figura 2.** Flux de treball MOCA

A continuació es mostren algunes pantalles de les principals interfícies: la d'introducció de metadades associades a un recurs; la de cerca que permet fer cerques combinades utilitzant qualsevol camp de metadades; i la d'actualització múltiple de registres.

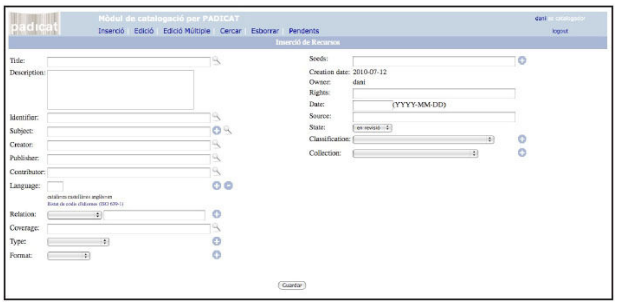

**Figura 3.** Interfície d'inserció de recursos MOCA v1.0

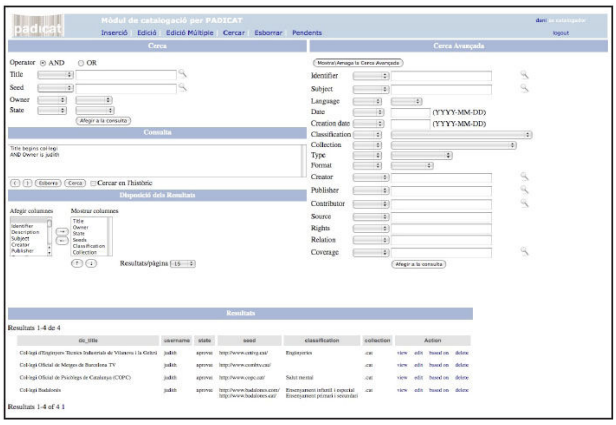

**Figura 4.** Interfície per la cera de recursos MOCA v1.0

Una vegada desenvolupada aquesta eina s'han creat uns scripts que han permès migrar la informació existent actualment de Web Curator Tool cap al nou model de dades de MOCA.

La data límit prevista per a la publicació en obert de la versió 1.0 de MOCA és a final de març de 2011.

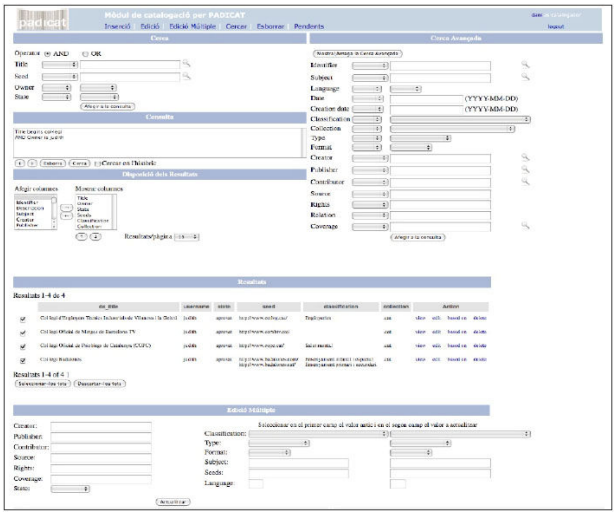

**Figura 5.** Interfície d'actualització múltiple MOCA v1.0

### 2.2. MOPU (Mòdul de Publicació)

MOPU permetrà generar automàticament el directori temàtic on estaran disponibles aquells recursos catalogats prèviament amb el MOCA que hagin estat seleccionats per a ser publicats.

Opcionalment cadascun d'aquests recursos podrà tenir associada una fitxa web, on es mostrarà informació descriptiva del recurs procedent de les metadades de catalogació que es considerin rellevants per a la visualització del recurs capturat. Les metadades que s'inclouen en aquesta fitxa són seleccionades de forma individualitzada i personalitzades per a cadascun dels recursos. És imprescindible disposar de MOCA per poder utilitzar MOPU.

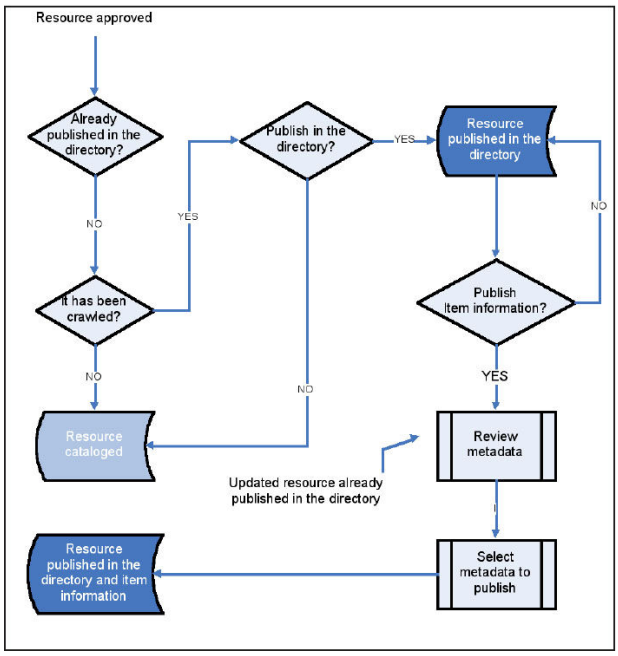


**Figura 6.** Flux de treball MOPU

Cal definir un nou rol per a MOPU, el validador. Aquest nou rol d'usuari serà l'encarregat de gestionar tot el procés de publicació. Comprovarà les dades introduïdes pel catalogador i donarà el vistiplau a la publicació del recurs en el directori temàtic i si s'escau quines metadades es visualitzaran a la fitxa web. En el cas que detecti alguna mancança en les metadades del recurs, el validador marcarà el recurs per a que el revisi el catalogador i a través d'un camp de notes associat al recurs comunicarà el motiu dels canvis a realitzar.

Paral·lelament es generarà de forma automàtica un històric de canvis que permetrà fer un seguiment sobre l'activitat de publicació i els canvis en la fitxa web del recurs.

A continuació es mostren algunes de les pantalles de les interfícies segons el pilot desenvolupat: la interfície de gestió del directori temàtic segons les categories; i la interfície per a la selecció de les metadades a incloure en la fitxa web d'un recurs.


**Figura 7.** Interfície del prototip MOPU


**Figura 8.** Interfície del prototip MOPU

### 2.3. MOST (Mòdul d'Estadístiques)

Aquest mòdul permetrà generar estadístiques sobre la informació inclosa en el repositori a partir dels reports generats pel programari de captura, essent exemples el nombre i la tipologia de fitxers que formen cadascuna de les captures preservades, volum i periodicitat de cada captura, etc. Aquest és un servei de valor afegit ja que des de PADICAT es poden consultar en obert tots els recursos capturats.

La integració d'aquest mòdul amb els anteriors permetrà incloure estadístiques per recurs en la fitxa web de presentació dels recursos preservats.

### 2.4. Full de ruta

Tan aviat com es disposi d'una versió estable i prou madura dels mòduls que formen CAT (Curator Archiving Tool), amb el feedback de la seva utilització, s'oferirà a la comunitat com a programari lliure. En el cas de MOCA està previst que sigui abans de 31 de març de 2011.

## 3. CONCLUSIONS

La captura de pàgines web publicades a Internet està suficientment resolta pels programaris ja madurs, com Heritrix. Per contra, els softwares d'indexació i els corresponents mòduls de visualització no han evolucionat al mateix ritme. Per arxius web com PADICAT, que ofereixen a Internet en obert la seva col·lecció, és imprescindible millorar els processos de cerca i visualització.

La concepció de l'eina CAT (Curator Archiving Tool) permet millorar la cerca i visualització dels recursos preservats en arxius web, gràcies a les seves aportacions en:

- La incidència humana en la descripció i indexació per mitjà del mòdul de Catalogació.
- La generació automatitzada de directoris i centres temàtics d'interès per mitjà del mòdul de Publicació.
- L'increment en l'oferta de dades estadístiques per mitjà del mòdul d'Estadístiques.

Per als gestors de l'arxiu web, la interacció dels tres mòduls permetrà beneficis en estalvi de temps derivat de l'automatització de processos actualment manuals:

- Integrar automàticament els recursos preservats en llistes alfabètiques, en directoris temàtics o en centres d'interès, que conformen l'accés a la col·lecció per mitjà de la navegació.
- Editar en grup recursos catalogats que comparteixin les mateixes característiques.
- Millorar els fluxos de treball entre els diferents rols que operen en la catalogació i publicació de recursos al portal d'accés a l'arxiu web.
- Adaptar PADICAT a l'estratègia de la Biblioteca Nacional de Catalunya, amb l'inici de la integració dels recursos preservats als catàlegs i sistemes de cerca ordinaris de la biblioteca.

Per als usuaris de l'arxiu web, la creació i posada en marxa de l'eina CAT (Curator Archiving Tool) permetrà una millora en la cerca i visualització dels recursos preservats:

- Cercar per navegació temàtica més eficaçment i exhaustivament, tot complementant els sistemes ja existents de cerca per paraula clau o bé per URL.
- Cercar per paraula clau més de manera pertinent, un cop s'intervingui en l'scoring dels sistemes de cerca, amb dades procedents de la catalogació.
- Cercar integradament, un cop s'integri la col·lecció de PADICAT als catàlegs i sistemes de cerca ordinaris de la Biblioteca de Catalunya.

- Accedir a informació de valor afegit per a cada recurs preservat, sigui procedent de la descripció humana dels recursos, sigui procedent de les dades estadístiques que crea el sistema de funcionament.

Per a la comunitat internacional, la publicació en obert dels mòduls de l'eina CAT (Curator Archiving Tool) permet:

- Contribuir a la missió d'adquirir, preservar i fer accessible la informació d'Internet per a futures generacions, arreu del món, promovent l'intercanvi global i les relacions internacionals.
- Potenciar la intervenció humana en els processos de descripció i publicació dels arxius web per millorar l'accés i visualització d'aquests sistemes.
- Aconseguir reconeixement públic i polític a la utilitat de projectes de preservació digital d'Internet, com són els arxius web protagonistes de la cita anual a l'IIPC *meeting*.