Facultat de Biblioteconomia i Documentació

Màster en Gestió de Continguts Digitals

Projecte TematiCAT

Autor: Joan Carles Pou Planas

Tutor: Miquel Térmens Graells

Juny de 2007

Agraïments

Aquest projecte ha estat per sobre de tot un repte. Res de tot el que he après i que he plasmat en les pàgines que segueixen hauria estat possible sense l'ajuda, suport, comprensió, fe, assessorament, etc., de tot un conjunt de gent que estat al meu voltant durant aquesta llarga etapa.

Primer de tot voldria agrair a Ciro Llueca per haver-me introduït en el món de la preservació digital i haver-me donat l'oportunitat d'experimentar amb un projecte tant important com el PADICAT.

També voldria donar les gràcies a tot el professorat del Màster en Gestió de Continguts Digitals que en el seu moment m'ha ajudat a trobar solucions impossibles a preguntes impossibles que jo els formulava. Especialment voldria agrair a Miquel Térmens tota la seva paciència i la bona predisposició que ha mostrat en tot moment per orientar-me dins el complex món de la gestió de projectes.

Finalment pels que han estat el dia a dia al meu voltant i han escoltat la frase ... *és que no veig clar el projecte*... *és que no tinc temps*... moltes gràcies per la vostra paciència, amistat i per tot el que heu fet per mi. Tot això no seria una realitat sense el vostre suport emocional. Gemma, Departament del CdD, gent de Figueres, gent d'Art, gent de Documentació, gent del Màster, gent del barri, de la ciutat i del món... gràcies de tot cor !

Sumari

1. Introducció

La situació general del català a Internet és bona[1]...

Aquesta cita, apareguda en un article publicat el mes de gener de 2007, argumenta que la presència del català en l'entorn web es consolida i que l'aparició de nous llocs web en aquesta llengua va *in crecendo*. Un estudi anterior fa una valoració de la salut del català a la xarxa a partir de xifres absolutes[2]. L'any 2005 es van comptar més de 7 milions de pàgines web, sent el català la 26a llengua amb més presència a Internet, i amb una mitjana de 1,09 pàgines web per parlant català, situant-se en el 19è lloc d'aquest *ranking*.

Per altra banda, el 16 de setembre de 2005 s'aprova el domini .cat, una nova iniciativa que ajudarà molt a millorar i consolidar la presència del català com una de les llengües amb més projecció a Internet, demostrant la força i reconeixement internacional d'aquesta llengua i cultura. Un any després de la data d'activació del domini .cat els dominis registrats ja superen els 21.000, i es preveu continuar amb un creixement regular[3].

Assolir aquestes xifres i l'alt grau d'implicació de la societat catalana ha requerit al llarg dels anys un gran esforç de conscienciació que la llengua i la cultura catalana també tenien lloc dins l'espai virtual del web. Ens hem de remuntar catorze anys enrere per conèixer la primera aparició del català a la xarxa, quan el fenomen web encara era molt minoritari i pràcticament només s'utilitzava la llegua anglesa. El primer lloc web en català va néixer a la Universitat Jaume I de la mà de Jordi Adell i els germans Carles i Antoni Bellver[4]. Es tractava d'un recurs molt senzill, composat per un logotip de la Universitat i uns enllaços que portaven a una pàgina de benvinguda, a un directori del personal de la institució i a una base de dades de la Biblioteca. Malgrat l'austeritat d'aquest recurs, tenint en compte l'òptica a la que estem acostumats en l'actualitat, va tenir el mèrit de ser un dels cent primers webs creats arreu del món. Ja només per

[1] SOLER MARTÍ, JOSEP. *Balanç de l´ús del català a Internet al 2006* [En línia]. Barcelona: Softcatalà, gener de 2007. [Data de consulta: 02/05/2007]. Disponible a: <http://www.softcatala.org/noticies/03012007510.htm>.

[2] MAS I HERNÁNDEZ, JORDI. *La salut del català a Internet el 2005* [En línia]. Barcelona: Softcatalà, octubre de 2005. [Data de consulta 02/05/2007]. Disponible a: <http://www.softcatala.org/articles/article60.htm>.

[3] PANTALEONI, ANA. "El dominio'.cat' obtiene 21.000 registros en un año" [En línia]. *El País.com Cataluña* Madrid, Prisacom S.A., gener de 2007. [Data de consulta: 02/05/2007]. Disponible a: <http://www.elpais.com/articulo/cataluna/dominio/cat/obtiene/21000/registros/ano/elpepuespcat/20070215elpcat_20/Tes>.

[4] PARTAL, VICENT. *El català a la xarxa* [En línia]*: història i raons d'un cas d'èxit*. Barcelona: Softcatalà, abril de 2004. [Data de consulta: 02/05/2007]. Disponible a: <http://www.softcatala.org/articles/article39.htm>.

aquest motiu es podria considerar el web de la Universitat Jaume I un petit tresor del patrimoni de la cultura catalana, però malauradament avui en dia ja no es conserva.

Relacionant les dades del nombre de webs que actualment existeixen i la història del primer web en català, ens podem imaginar la quantitat d'informació que pot arribar a desaparèixer, ja sigui perquè els creadors no la conservin, o bé perquè l'evolució dels formats web facin impossible la lectura d'aquests. Aquesta informació susceptible de perdre's forma part del Patrimoni Bibliogràfic Digital Català, i de la mateixa manera que avui lamentem la pèrdua de la primera manifestació catalana a la xarxa (entre molts altres), podríem lamentar en un futur proper la desaparició d'un nombre incomptable de webs.

Com a reflex de diversos projectes portats a terme arreu del món i amb l'ànim de poder aportar una solució a la problemàtica que es planteja, l'any 2005 neix una iniciativa en el si de la Biblioteca de Catalunya, el projecte PADICAT, amb l'objectiu de preservar el patrimoni digital català.

La missió del projecte PADICAT és dissenyar i produir un sistema que permeti a la Biblioteca de Catalunya compilar, processar i donar accés **permanent** a part de la producció digital catalana. Aquest projecte pren la forma d'un dipòsit de recursos digitals definit pel desenvolupament de tres objectius estratègics:

- Compilació massiva de recursos publicats en obert a Internet
- Impulsió d'un dipòsit sistemàtic de la producció web dels agents implicats a Catalunya
- Promoció de línies de recerca per mitjà de la integració dels recursos digitals de determinats esdeveniments de la vida pública catalana

Des de l'octubre de 2006 el projecte PADICAT ja està en fase operativa i és accessible en obert a la xarxa. Tanmateix, s'ha fixat l'any 2009 com a data de finalització del desenvolupament del projecte i consolidació del seus sistemes.

El projecte que es presenta en aquesta memòria està basat en el desenvolupament del tercer objectiu estratègic del PADICAT, que implica la creació d'un sistema tecnològic de captura, tractament i accés de recursos temàtics relacionats amb esdeveniments de la vida pública catalana.

1.1. Motivació del projecte

A patir d'unes reunions informals mantingudes durant l'any 2006 amb Ciro Llueca, coordinador del projecte PADICAT, i en relació a la realització d'un projecte en el marc del Màster en Gestió de Continguts Digitals, es va plantejar la possibilitat de treballar en el desenvolupament d'algun dels aspectes del PADICAT. La proposta de desenvolupament del projecte es va concretar en el disseny d'un sistema tecnològic que permetés la captura, el tractament i l'accés d'una sèrie de recursos temàtics relacionats amb esdeveniments de la vida pública catalana.

Davant la dificultat d'emprendre aquest projecte, degut a la manca d'experiència i a la impossibilitat de poder tenir accés a certa informació[5], es va acordar amb Ciro Llueca i amb el tutor del projecte TematiCAT, Miquel Térmens, que la planificació d'aquest es realitzaria des d'una vessant purament acadèmica, tenint com a finalitat última l'adquisició de nous coneixement i la defensa d'un projecte per obtenir la titulació del Màster de Gestió de Continguts Digitals.

Degut a la contínua evolució del projecte PADICAT, amb el que està directament relacionat el projecte TematiCAT, es va prendre com font d'informació principal la *Memòria del plantejament del projecte PADICAT*[6], presentada el desembre de 2005. Això implica que els canvis i transformacions sofertes en el projecte PADICAT posteriorment a la presentació de la memòria, no es reflectiran en el desenvolupament del projecte TematiCAT. Com excepció d'aquesta premissa, s'ha tingut en compte la pàgina web del PADICAT com objecte d'anàlisi de la interfície de cerca de l'usuari final, ja que permet extreure una sèrie d'informació que no es troba en la *Memòria* de desembre de 2005.

En el terreny pròpiament personal, la motivació per portar a terme aquest projecte ve donada per la inquietud al voltant de l'entorn web i tots els aspectes que fan referència a la publicació d'informació en format digital a la xarxa i a la seva preservació. Altres aspectes com l'admiració per projectes com Internet Archive o Minerva han contribuït en l'ànim de treballar en un projecte com el TematiCAT.

[5] Al llarg del desenvolupament del projecte TematiCAT s'han mantingut reunions amb Ciro Llueca per tal de poder recavar informació necessària per treballar alguns aspectes més tècnics del projecte. En la majoria d'ocasions aquesta necessitat d'informació es va poder satisfer, però en altres casos va ser impossible poder respondre certes preguntes, degut a que eren problemàtiques que encara no s'havien abordat en el projecte PADICAT.

[6] BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)* [En línia]. Barcelona: Biblioteca de Catalunya, desembre 2005. [Data de consulta: 09/05/2007]. Disponible a: <http://www.recercat.net/handle/2072/1757>.

1.2. Metodologia

En la planificació d'aquest projecte s'ha seguit una aproximació metodològica per assolir els objectius marcats, generant un pla de treball basat en l'equilibri entre la inversió de temps i recursos i la consecució de resultats. Les principals etapes en que s'ha subdividit el desenvolupament del projecte són les següents:

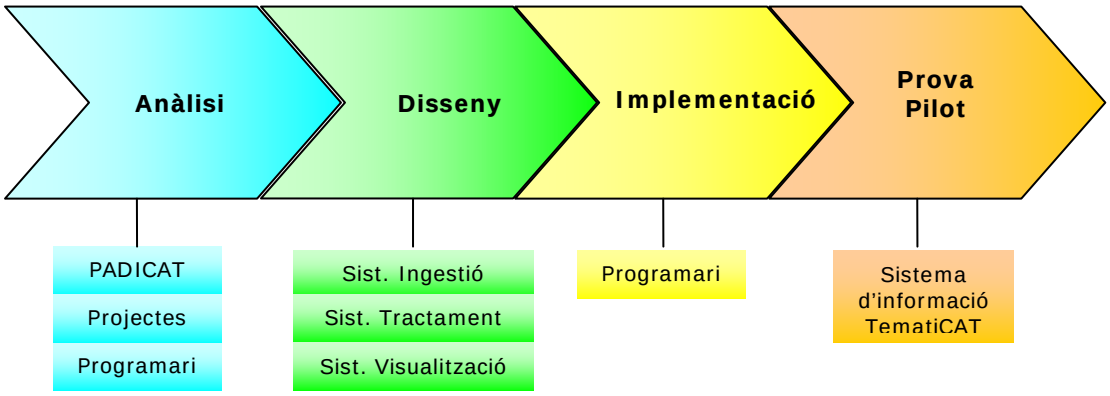


Fig. 1. Aproximació metodològica de la planificació del projecte TematiCAT

- ***Anàlisi:*** la fase d'anàlisi representa un gruix important del projecte perquè s'ha hagut de documentar un entorn que ha evolucionat molt ràpidament amb pocs anys, i la seva complexitat continua en procés de creixement. L'anàlisi realitzat en aquest projecte s'ha basat en quatre pilars fonamentals: anàlisi de l'abast i funcionament del projecte PADICAT, anàlisi del context de la preservació digital, anàlisi de projectes similars a nivell internacional i anàlisi dels recursos tecnològics existents[7]. Fruit d'aquesta recopilació d'informació s'ha generat un coneixement que s'ha aplicat en l'estudi i disseny del sistema d'informació que gestionarà tots els recursos del TematiCAT.
- ***Disseny:*** per tal de reduir la complexitat de la construcció del sistema d'informació del TematiCAT, s'ha subdividit el desenvolupament del sistema en tres mòduls funcionals en base dels processos que realitzen (selecció i captura, tractament i accés). El disseny dels mòduls del sistema s'han realitzat en base dels exemples observats en altres projectes i sempre segons els objectius marcats per l'abast del projecte.

[7] L'anàlisi dels recursos tecnològics, tot i tractar-se d'un contingut propi d'aquesta primera fase de la metodologia que s'exposa, ha estat introduït en l'apartat del disseny del sistema tecnològic del projecte, per tal relacionar-lo amb els requisits que s'exposen en cada mòdul del sistema.

- ***Implementació:*** en base el prototip del sistema d'informació dissenyat i les característiques dels recursos tecnològics recollits en l'anàlisi del programari, s'han escollit unes aplicacions que s'encarregaran de dur a terme el diferents processos de la plataforma. En aquesta fase s'ha plasmat un pla de treball on es determinen les parametritzacions que s'hauran de realitzar per adaptar el funcionament del programari als requisits del projecte. En aquesta fase també s'han programat una sèrie de proves pilot destinades a comprovar l'èxit de la implementació de cada mòdul per tal de poder ajustar la seva configuració.
- ***Prova Pilot:*** en aquesta fase es detallarà la simulació de la captura d'un esdeveniment amb la finalitat de comprovar el funcionament del sistema d'informació de forma global, i també com exemple de com procedimentar les tasques que s'hauran de realitzar.

2. Context del projecte TematiCAT

La missió

Crear una solució tecnològica que sistematitzi la captura de recursos digitals en línia d'interès patrimonial català segon criteris temàtics, i la seva posterior validació i catalogació, amb la finalitat última de fer-los accessibles a la societat de forma perpètua assegurant la seva preservació.

La visió

Ser un projecte de referència en l'entorn d'Internet i dins el panorama cultural català. En aquest sentit és important que es percebi el projecte com una plataforma que s'interessa per la conservació de la producció de recursos web catalans, per tal de motivar a altres institucions a col·laborar en la preservació d'aquest Patrimoni.

El nom

La denominació que rep el projecte, TematiCAT ve donada per la fusió dels conceptes *temàtica* i *Catalunya*, ja que es tracta d'una web que contindrà col·leccions de recursos temàtics relacionats amb esdeveniments a Catalunya. A l'hora de donar nom al projecte s'ha valorat la significació, la sonoritat i la simplicitat amb la finalitat que sigui fàcil de recordar. Les lletres CAT es ressalten en majúscula conferint un cert simbolisme al nom del projecte, ja que d'aquesta manera es remarca la vinculació amb Catalunya, i alhora es relaciona amb la terminació del nom PADICAT. També és important emfatitzar que TematiCAT s'afegeix a la llista d'una sèrie de projectes que utilitzen el recurs CAT en el seu nom, com són TERMCAT, RECERCAT i el mateix PADICAT.

L'organització

El PADICAT, projecte sobre el que treballarà el TematiCAT, es desenvolupa en el si de la Biblioteca de Catalunya, sent aquesta una institució pública que depèn de la Generalitat de Catalunya (*Annex 1*). La Biblioteca de Catalunya, en tant que Biblioteca Nacional, desenvolupa la tasca de dipòsit legal a nivell català, i és dins aquest marc on es relaciona el projecte de preservació de la producció bibliogràfica de Catalunya.

2.1. Abast de projecte

A qui beneficia el projecte?

El projecte pretén tenir dos tipus de beneficiaris. El primer, i de forma directe, són els usuaris que facin ús del TematiCAT consultant els recursos als quals s'ofereix accés. El segon, més abstracte i de forma indirecte, és la cultura catalana, ja que aquest projecte col·labora en la difusió del patrimoni digital català i n'assegura la seva preservació.

Qui són els nostres usuaris?

El nostre usuari potencial pot ser qualsevol persona interessada en la informació relacionada amb l'àmbit català que es pot trobar en l'entorn d'Internet. Així doncs, es tractarà d'usuaris heterogenis, sense un perfil concret, caracteritzats únicament per les inquietuds informatives a nivell de Catalunya.

Abast temàtic i temporal

L'abast temàtic del projecte estarà centrat en la recopilació de recursos digitals que informen sobre certs esdeveniments d'interès general que es produeixen a Catalunya. Aquest abast és el principal criteri que delimita el radi d'acció del projecte i que alhora en justifica la seva existència. Les diferents col·leccions temàtiques generaran una instantània que cobrirà la duració dels esdeveniments i les seves repercussions immediates, amb la finalitat de poder copsar de forma global gran part del que s'ha publicat a la xarxa sobre un determinat fet.

Abast geogràfic

El projecte és una iniciativa que es vincula amb la missió de conservació del patrimoni bibliogràfic català de la Biblioteca Nacional de Catalunya. Com a tal, el projecte es centrarà en l'àmbit geogràfic de Catalunya.

2.2. Objectius estratègics

- ***Analitzar l'abast i el funcionament del projecte PADICAT***: s'haurà de realitzar un estudi profund dels objectius i de tota l'estructura sobre la que treballa el sistema tecnològic del PADICAT, perquè la plataforma del TematiCAT hi haurà d'interactuar. Per la relació que hi haurà entre els dos projectes és molt important conèixer minuciosament els seus detalls

- ***Analitzar els casos de projectes similars a nivell internacional***: ser innovador quan es parla de tecnologia web és arriscat. Per aquest motiu es realitzarà un ampli anàlisi de projectes que realitzen funcions similars al TematiCAT, amb la finalitat de conèixer el seu funcionament, posant èmfasi en els èxits i fracassos obtinguts.

- ***Dissenyar i implementar un sistema de selecció i captura temàtica de recursos***: creació d'un sistema tecnològic amb l'objectiu d'aconseguir la màxima automatització possible dels processos de selecció i captura de recursos temàtics, reduint el temps i els costos, vers el que seria una selecció manual, i mantenint un alt nivell d'encert en la captura.

- ***Dissenyar i implementar un sistema de tractament dels recursos***: creació d'un sistema tecnològic que permeti la validació, indexació i descripció dels recursos amb l'objectiu de poder ser recuperats posteriorment.

- **Dissenyar i implementar un sistema de recuperació i visualització als recursos temàtics**: creació d'un sistema tecnològic que possibiliti als usuaris la recuperació dels recursos web allotjats en el dipòsit, mitjançant l'ús de diferents eines de cerca.

- **Realització d'una col·lecció temàtica BETA, basada en les eleccions municipals de 2007**: la creació d'aquesta col·lecció temàtica té com objectiu la realització d'un prova pilot global de tot el sistema tecnològic del TematiCAT.

En aquest marc dels objectius estratègics del projecte es vol reiterar que la missió es centra en la creació d'un sistema tecnològic que engloba els processos de selecció, captura, tractament i accés de recursos temàtics relacionats amb esdeveniments. En aquest sentit es vol posar èmfasi en que no es contempla en la planificació l'objectiu de fer operatiu el projecte. Per aquest motiu, i alhora per centrar el projecte en el

desenvolupament d'una solució tecnològica, no s'han abordats objectius com la implementació de la interfície, ja que els aspectes d'usabilitat, accessibilitat i disseny gràfic podrien representar tot un nou projecte, ni tampoc la difusió del projecte i la posterior avaluació.

2.3. Anàlisi DAFO

Punts forts

-El projecte s'emmarca en un programa més ampli ja consolidat

-El projecte treballarà en base recursos web que ja hauran estat estudiats en fases anteriors

-Existeixen projectes semblants arreu del món que s'han culminat amb èxit i que serviran de referent

Punts febles

-Manca d'experiència en el terreny de la gestió de recursos digitals

-Costos molt importants en inversió tecnològica i la seva implementació

-Ús d'aplicacions tecnològiques no estables pel seu continu desenvolupament

Oportunitats

-L'alliberament del domini punt cat (.cat) facilitarà el treball de recopilació

-Recolzament total per part de la direcció de la Biblioteca de Catalunya

-Moment conjuntural polític que afavoreix la inversió en els aspectes culturals catalans

-Internet cada dia està més arrelat a la societat i està sent una de les vies més importants d'accés a la informació

-Possibilitat d'esdevenir un model per futurs projectes

Amenaces

-Possibilitat d'un canvi de govern en les pròximes eleccions que pogués modificar els objectius estratègics de la Biblioteca de Catalunya

-Desconeixement de com pot evolucionar la tecnologia web i quines repercussions pròximes hi pot haver respecte el projecte

-Existència moltes llacunes en l'àmbit legal del tractament de recursos digitals

2.4. Requeriments de recursos humans

En el projecte hi participaran quatre professionals: el cap de projecte, un tècnic de sistemes i dos documentalistes, dels qual tots seguit se'n detalla el seu corresponent perfil:

Cap de projecte	
Acadèmic	Llicenciatura en Documentació
Formació addicional	Màster en gestió de continguts digitals
Professional	Experiència en gestió de recursos web
Competències genèriques	Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Iniciativa Flexibilitat Sentit analític i crític
Competències professionals	Comunicació Anglès Negociació Visió Direcció i gestió Legislació Administració Lideratge Presa de decisions
Fases de participació	Anàlisi Disseny Implementació Col·lecció temàtica Beta

Tècnic de sistemes	
Acadèmic	Enginyeria o Informàtica superior
Formació addicional	Especialització en TIC
Professional	Experiència en projectes similars
Competències genèriques	Treball en equip Relacions interpersonals Orientació als resultats Planificació i organització Iniciativa Flexibilitat Sentit analític i crític Documentació de processos
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
Fases de participació	Anàlisi Disseny Implementació Col·lecció temàtica Beta

Documentalista	
Acadèmic	Diplomatura en Biblioteconomia o Llicenciatura en Documentació
Formació addicional	Cursos d'anàlisi i descripció de documents
Professional	Experiència en descripció de recursos
Competències genèriques	Treball en equip Orientació als resultats Planificació i organització Sentit analític i crític
Competències professionals	Gestió de la col·lecció Gestió de continguts Anàlisi i representació documental Descripció i organització documental Recuperació de la informació
Fases de participació	Col·lecció temàtica Beta

2.5. Requeriments tecnològics

A nivell de programari, aquest projecte apostarà totalment per les aplicacions en codi obert. El sistema operatiu serà Linux, i Apache 2 i Tomcat 5 seran les aplicacions de servidor. Pel que fa als programes específics per a cada procés, s'instal·laran les següents aplicacions que seran extensament tractades en els apartats dedicats al disseny i implementació:

- ***Heritrix***: programa de captura de recursos
- ***Nutch Wax***: programa d'indexació i motor de cerca
- ***Web Curator***: programa d'administració dels
- ***Wera***: programa de recuperació i visualització

Pel que fa al maquinari, els components que faran operatiu el sistema d'informació del TematiCAT són els següents:

- 2 Servidors SUN Fire V490
- Dipòsit d'emmagatzemament (4 Tb) *Sun StorageTek 3510 FC Array*
- Llibreria de cintes LTO (4 Tb) Sun StorageTek C4
- Servidor de pàgina web IBM xseries 346 (8840ecg)
- 3 Ordinadors (2 Gb. memòria; processador 5.2 GHz; 320 Gb HD)
- Impressora HP DeskJet 9800
- 3 SAIS APC Smart-UPC RT 3000 VA

2.6. Definició de les tasques del projecte

Tasques	Responsable	Participants	Duració
Anàlisi de l'abast i funcionalitats del projecte			
Reunió amb el coordinador del projecte PADICAT	Cap de projecte TematiCAT	Coordinador PADICAT	4 hores
		Cap de projecte TematiCAT	4 hores
Definició de l'abast del projecte	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
Definició dels objectius del projecte	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
Estudi dels requeriments de recursos humans, infraestructures i materials	Cap de projecte TematiCAT	Cap de projecte TematiCAT	6 hores
Estudi dels requeriments tecnològics	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
		Informàtic	12 hores
Establiment d'un timing de la planificació i disseny del projecte	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
		Informàtic	8 hores
Anàlisi de l'abast i funcionament del projecte PADICAT			
Estudi de la memòria de plantejament publicada l'any 2005	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
Reunions amb el coordinador del projecte PADICAT	Cap de projecte TematiCAT	Coordinador PADICAT	8 hores
		Cap de projecte TematiCAT	8 hores
Anàlisi de les característiques tecnològiques del projecte (maquinari, programari, sistemes operatiu, llenguatges de programació, etc.)	Cap de projecte TematiCAT	Cap de projecte TematiCAT	15 hores
		Informàtic	15 hores
Anàlisi de les característiques dels recursos web capturats (nombre, idioma, format, volum, etc.)	Cap de projecte TematiCAT	Cap de projecte TematiCAT	8 hores
Anàlisi dels processos del projecte • Captura (automàtica i per acords) • Emmagatzematge • Tractament (automàtic i manual) • Recuperació de recursos • Visualització	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	20 hores
Anàlisi de les necessitats i tasques realitzades pels recursos humans	Cap de projecte TematiCAT	Cap de projecte TematiCAT	6 hores
Anàlisi del pressupost del projecte	Cap de projecte TematiCAT	Cap de projecte TematiCAT	4 hores
Redacció d'un informe de l'anàlisi realitzat	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores

Tasques	Responsable	Participants	Duració
Anàlisi de casos de projectes similars a nivell internacional			
Recopilació i lectura bibliografia sobre el context dels dipòsits digitals a nivell mundial	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
Recopilació i lectura bibliografia sobre projectes similars	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
Elaboració una llista d'indicadors per poder comparar els diferents projectes	Cap de projecte TematiCAT	Cap de projecte TematiCAT	4 hores
Elaboració d'un estudi de 6 projectes que porten a terme la selecció d'esdeveniments temàtics, i valorar els aspectes d'èxit i fracàs	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
Elaboració d'un balanç dels factors crítics d'èxit dels projectes que es poden incorporar al TematiCAT, i les problemàtiques a evitar	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
Disseny del sistema de selecció i captura temàtica de recursos digitals			
Definició de les funcionalitat del sistema de selecció i captura	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	16 hores
Definició de les estratègies de captura selectiva de recursos digitals	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
Creació d'un patró de les característiques dels recursos digitals	Cap de projecte TematiCAT	Cap de projecte TematiCAT	15 hores
Definició dels requeriments tecnològics del programa que ha de realitzar el procés de selecció de recursos del dipòsit PADICAT i la indexació	Cap de projecte TematiCAT	Cap de projecte TematiCAT	8 hores
		Informàtic	8 hores
Estudi de diferents programes de selecció i indexació que compleixen els requeriments tecnològics	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	20 hores
Definició dels requeriments tecnològics del programa que ha de realitzar el procés de captura de recursos a Internet	Cap de projecte TematiCAT	Cap de projecte TematiCAT	8 hores
		Informàtic	8 hores
Estudi de diferents programes de captura que compleixen els requeriments tecnològics	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	20 hores
Disseny del sistema de tractament dels recursos digitals			
Definició de les funcionalitat del sistema de validació i descripció	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	14 hores
Definició dels requeriments tecnològics del programa que ha de realitzar el procés de validació i descripció	Cap de projecte TematiCAT	Cap de projecte TematiCAT	8 hores
		Informàtic	8 hores
Estudi de diferents programes que compleixen els requeriments tecnològics	Cap de projecte TematiCAT	Cap de projecte TematiCAT	10 hores
		Informàtic	10 hores

Tasques	Responsable	Participants	Duració
Disseny del sistema de recuperació i visualització dels recursos digitals capturats			
Definició de les funcionalitat del sistema de recuperació i visualització	Cap de projecte TematiCAT	Cap de projecte TematiCAT	20 hores
		Informàtic	16 hores
Definició dels requeriments tecnològics del programa que permeti la recuperació i visualització del recursos	Cap de projecte TematiCAT	Cap de projecte TematiCAT	12 hores
		Informàtic	12 hores
Estudi de diferents programes que compleixen els requeriments tecnològics	Cap de projecte TematiCAT	Cap de projecte TematiCAT	15 hores
		Informàtic	15 hores
Disseny d'un prototip d'interfície de consulta per a l'usuari final	Cap de projecte TematiCAT	Cap de projecte TematiCAT	8 hores
Implementació del programari			
Implementació del Sistema Operatiu i els programes Apache 2 i Tomcat 5	Informàtic	Informàtic	16 hores
Implementació i prova pilot de l'aplicació Heritrix	Informàtic	Cap de projecte TematiCAT	30 hores
		Informàtic	70 hores
Implementació i prova pilot de l'aplicació Nutch Wax	Informàtic	Cap de projecte TematiCAT	30 hores
		Informàtic	70 hores
Implementació i prova pilot de l'aplicació Web Curator	Informàtic	Cap de projecte TematiCAT	30 hores
		Informàtic	60 hores
Implementació i prova pilot de l'aplicació Wera	Informàtic	Cap de projecte TematiCAT	30 hores
		Informàtic	60 hores
Creació d'una col·lecció temàtica Beta			
Planificació i abast de la col·lecció	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
Preselecció de recursos	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
		Documentalistes	50 hores
Captura de recursos	Cap de projecte TematiCAT	Cap de projecte TematiCAT	40 hores
		Informàtic	20 hores
Tractament dels recursos capturats	Documentalistes	Cap de projecte TematiCAT	50 hores
		Documentalistes	1100 hores
Proves pilot del sistema d'accés als recursos	Cap de projecte TematiCAT	Cap de projecte TematiCAT	30 hores
		Informàtic	30 hores
		Documentalistes	30 hores
Avaluació i modificacions del sistema	Cap de projecte TematiCAT	Cap de projecte TematiCAT	40 hores
		Informàtic	40 hores

2.7. Timing

La planificació del projecte tindrà una duració de 10 mesos, amb data d'inici a l'octubre de 2006 i de finalització el juliol de 2007. En el següent quadre es desglossa el període d'activitat del projecte TematiCAT.

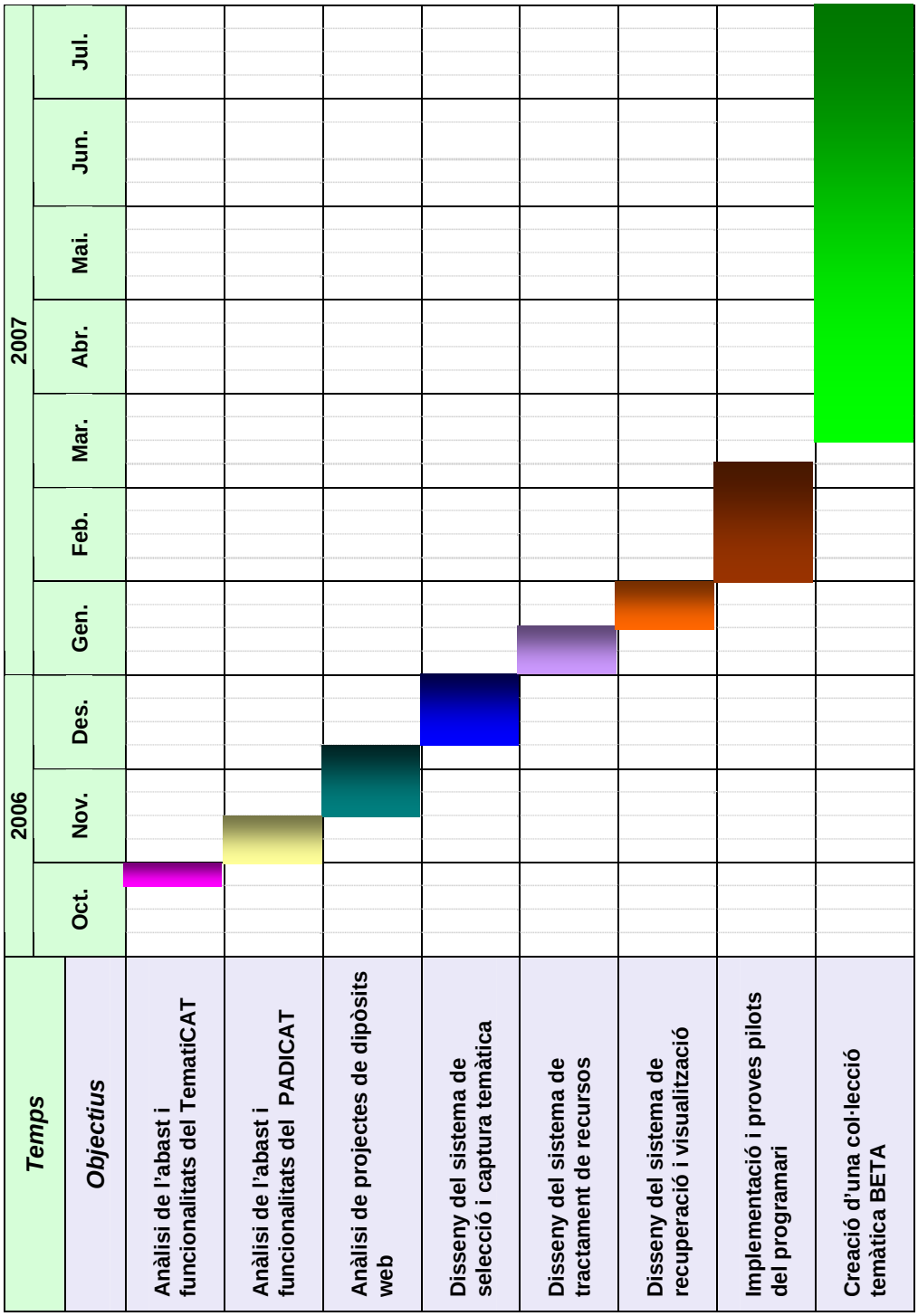


Fig. 2. Timing de projecte TematiCAT

2.8. Pressupost

El pressupost previst pel desenvolupament del projecte TematiCAT és de 167.558,9 €, i inclou les següents despeses:

Recursos humans	55.328,0 €
Material tecnològic	74.036,3 €
Mobiliari	4.960,0 €
Despeses generals	26.790,0 €
Subtotal	**161.144,3 €**
+ 4% per contingències	6.444,6 €
Total	**167.558,9 €**

Fig. 3. Pressupost total del projecte

En les següents taules es desglossa cada apartat del pressupost amb els seus corresponents detalls.

Recursos Humans			
Professional	***€ per hora (bruts)***	***Nombre d'hores***	***Total***
Cap de projecte	22 €	791	22.848,8 €
Tècnic de sistemes	18 €	556	13.140,5 €
Documentalista	12 €	1180	18.592,0 €
Coordinador PADICAT	47,4 €	12	746,7 €
			55.328,0 €

Fig. 4. Pressupost en recursos humans

La remuneració en brut de cada perfil professional inclou el percentatge corresponent d'IRPF i seguretat social, i al total se li ha sumat la quota patronal (31,3%), obtenint per tant el cost empresarial de cada participant del projecte[8].

[8] Les remuneracions han estat extretes del Manual de remuneració de Ceinsa. **Vegeu**: GRUPO RH ASESORES. *Informe de remuneración España* 2006-2007 [Recurs electrònic]. Madrid: Ceinsa, 2006. En el cas de la participació del Coordinador del PADICAT se li ha aplicat una tarifa d'assessorament. **Vegeu**: COL·LEGI OFICIAL DE BIBLIOTECARIS I DOCUMENTALISTES DE CATALUNYA. *Quant cobra un bibliotecari-documentalista?* [En línia]. Barcelona, COBDC, abril de 2007. [Data de consulta 23/05/2007]. Disponible a: <http://www.cobdc.org/serveis/assessoria.html#01>.

Material tecnològic			
Component	***Preu unitat***	***Unitats***	***Total***
Servidor SUN Fire V490	23.020,6 €	2	46.041,2 €
Dipòsit d'emmagatzemament (4 Tb) *Sun StorageTek 3510 FC Array*	7052,1 €	1	7052,1 €
Llibreria de cintes LTO (4 Tb) Sun StorageTek C4	7.427,2 €	1	7.427,2 €
Servidor per pàgina web IBM xseries 346 (8840ecg)	3808.8 €	1	3808,8 €
Cables	1.500 €	---	1.500,0 €
Ordinador	900 €	3	2.700,0 €
Impressora HP DESKJET 9800 30PPM	314€	1	314,0 €
SAI *APC Smart-UPS RT 3000VA*	1731 €	3	5193,0 €
			74.036,3 €

Fig. 5. Pressupost en material tecnològic

El pressupost del material s'ha basat en els costos fixes de propietat i no s'han tingut en compte les amortitzacions dels components en base a la duració del projecte. El pressupostos del components a gran escala, servidors i dipòsits d'emmagatzematge s'han obtingut dels distribuïdors de les marques SUN[9] i IBM[10], mentre que la resta components s'han pressupostat a la cadena de botigues d'informàtica Pricoinsa[11].

Mobiliari			
Producte	***Preu Unitat***	***Nombre d'hores***	***Total***
Taules	300 €	4	1.200,0 €
Cadires	110 €	6	660,0 €
Il·luminació	1.000 €	---	1.000,0 €
Prestatgeries	200 €	2	400,0 €
Arxivadors	200 €	1	200,0 **€**
Mobiliari informàtic	1.500 €	1	1.500,0 €
			4.960,0 €

Fig. 6. Pressupost en mobiliari

[9] SUN MICROSYSTEMS. *Productos y servicios* [En línia]. Madrid: Sunmicrosystems, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://es.sun.com/productos_servicios/>.

[10] IBM. *IBM Servers* [En línia]. New York: IBM, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://www-03.ibm.com/servers/>.

[11] PRICOINSA. *Bienvenido a Pricoinsa* [En línia]. Barcelona: Pricoinsa, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://www.pricoinsa.es/>.

Despeses Generals			
Producte	***Preu per mes***	***Mesos***	***Total***
Fungibles	100 €	10	1.000,0 €
Connexió Internet	100 €	12	1.200,0 €
Correu electrònic	30 €	---	30,0 €
Domini Web	60 €	---	60,0 €
Consums (llum, tel. aigua, etc.)	350 €	10	3.500,0 **€**
Lloguer Oficina 200 m^2	2.000 €	10	20.000,0 €
Despeses d'administració	1.000 €		1.000,0 €
			26.790,0 €

Fig. 7. Pressupost de les despeses generals

En aquest últim apartat s'han agrupat les despeses de consum generals i d'administració, així com el cost del lloguer de l'espai on es realitzarà el projecte.

3. Fase d'anàlisi

La fase d'anàlisi té un pes molt important en la planificació del projecte, ja que en aquesta es recavarà tota la informació necessària per dissenyar un sistema d'informació capaç de gestionar tots els processos que haurà de realitzar la plataforma tecnològica del TematiCAT. En aquest apartat es farà un retrat superficial de l'estat de la qüestió de la conservació del patrimoni digital, i tot seguit s'introduirà un anàlisi realitzat a 6 projectes que estan actuant en l'òrbita de la preservació dels recursos web. L'últim punt d'aquest apartat es dedicarà a l'anàlisi de l'abast i funcionalitats del projecte PADICAT, en base a la *Memòria del plantejament del projecte* de 2005, el qual ha de servir de base pel desenvolupament del projecte que es presenta.

3.1. Estat de la qüestió global del patrimoni digital

Cada vegada més el nostre patrimoni cultural, científic i de la informació està prenent forma digital, ja sigui perquè es digitalitzen els formats analògics tradicionals, o bé perquè són creats directament en format digital (*born digital*).

El progrés accelerat de les tecnologies de la informació i la comunicació estan molt relacionades amb l'apogeu del procés de digitalització d'aquest patrimoni i de les noves formes de transmissió d'aquest a gran part del món.

Davant d'aquesta sobtada transformació del patrimoni cultural mundial, la UNESCO va publicar les *Directrices para la preservación del patrimonio digital*[12], amb la voluntat de conscienciar els productors i gestors d'aquest patrimoni, i vetllar per la seva conservació. Les Directrius reflecteixen que els objectes que formen el patrimoni digital poden ser des de textos, bases de dades, imatges fixes o en moviment, gravacions sonores, material gràfic, programes informàtics o pàgines web[13], entre molts altres formats possibles dins un ampli repertori de diversitat creixent.

La UNESCO posa en relleu que molts d'aquests recursos posseeixen una gran importància i valor durador, i constitueixen un patrimoni digne de ser protegit i conservat en benefici de les generacions actuals i futures. Per aquest motiu, l'objectiu

[12] Biblioteca Nacional de Austràlia. *Directrices para la preservación del patrimonio digital* [En línia]. Canberra : Unesco, 2003. [Data de consulta: 05/05/2006]. Disponible a: <http://unesdoc.unesco.org /images/0013/001300/130071s.pdf>.

[13] Entendrem el concepte pàgina web en el sentit més ampli, incloent portals, llocs web, blogs, etc.

de la conservació del patrimoni digital és que aquest sigui accessible per la societat de forma permanent, de la mateixa manera que ho estan els que s'han publicat en formats tradicionals.

Comparant actualment la progressió de la producció digital davant de les pròpies mesures de conservació, es fa palès que el patrimoni digital mundial està en perill per la posteritat. Hi ha molts factors que contribueixen a fer creixent aquesta amenaça, com la ràpida obsolescència dels equips i programes informàtics que donen vida als objectes digitals, les incerteses existents al voltant dels recursos, la responsabilitat i els mètodes pel seu manteniment i conservació, i sobretot la manca d'un marc legal que empari aquests processos, tal i com s'ha fet al llarg de la història amb el patrimoni cultural tradicional.

L'evolució tecnològica ha estat tant ràpida, que governs i institucions no han pogut elaborar estratègies de conservació oportunes i ben fonamentades. La inestabilitat d'aquests recursos digitals posa en perill el patrimoni intel·lectual i cultural de la humanitat, i per extensió l'equilibri econòmico-social. S'ha arribat a un punt que es fa imprescindible començar a emprendre accions de divulgació i promoció sobre els responsables de formular polítiques, i sensibilitzar el públic en general sobre el potencial del recursos digitals i dels problemes pràctics que planteja la seva preservació.

La UNESCO posa en relleu que els poders públics han de prendre la iniciativa davant el panorama actual del mercat lliure, en el que preval el benefici econòmic immediat i la innovació constant per sobre de la creació d'un context que faciliti la conservació dels objectes digitals.

Estats i institucions han d'intercedir per crear mecanismes adequats per garantir la perdurabilitat del patrimoni digital. Fer que la legislació de dipòsit legal o voluntari a les biblioteques, arxius, museus o altres institucions públiques de conservació s'apliqui al patrimoni digital ha de ser un element prioritari a la política nacional de preservació. En aquest aspecte, un punt clau és la legislació sobre drets d'autor i drets connexos, la qual s'ha de reelaborar tenint en compte totes les noves implicacions des del punt de vista del recursos digitals, per tal de permetre i facilitar que les esmentades institucions puguin portar a terme el procés de conservació digital.

3.1.1. Evolució de les accions de preservació digital

Dins el context que s'ha esbossat en l'apartat anterior, l'objectiu de conservar la producció digital obeeix una evolució lògica. Inicialment les primeres manifestacions de voluntat de preservació digital les trobem en les biblioteques de recerca, universitàries, nacionals i també públiques, que per mitjà de les biblioteques virtuals elaboraven directoris de recursos electrònics[14].

El següent pas, amb la missió de complementar les necessitats temàtiques concretes, va ser la creació de dipòsits[15] multiformats, que incloïen imatge, so, text, gràfics, etc. Aquestes recopilacions multimèdia s'organitzaven habitualment sota criteri geogràfic i seguien el model enciclopèdic digital del CD-ROM.

El tercer pas d'aquesta evolució dins la preservació digital va ser la creació de dipòsits institucionals, els quals tenien la missió de conservar la producció digital d'una organització. Aquesta tipologia de dipòsits són molt habituals en institucions d'investigació i recerca, perquè els permet conservar el fruit de la seva activitat, i alhora, els ofereix una plataforma de treball col·laboratiu i divulgació externa de la seva tasca.

Actualment ja podem parlar d'una fase avançada d'una altra generació de dipòsits digitals, els nacionals. Aquests dipòsits estan liderats per les biblioteques nacionals d'alguns països i tenen la missió de garantir l'accés a llarg termini als recursos digitals que es generen en un territori, o sobre un territori determinat. Concretant en aquest àmbit, la missió de la Biblioteca Nacional de Catalunya, en tant que biblioteca nacional, és la de *recollir, conservar i difondre la producció bibliogràfica catalana i la relacionada amb l'àmbit lingüístic català, i vetllar per la conservació i la difusió del patrimoni bibliogràfic*[16]*.* Aquesta premissa és la que s'ha utilitzat tradicionalment per la producció impresa catalana, emparada per la llei de dipòsit legal, i que avui en dia ha hagut d'eixamplar els seus horitzons, a causa del context que estem vivint, per tal d'incloure la producció bibliogràfica digital.

---

[14] LLUECA FONOLLOSA, CIRO. "Webs sempre accessibles [En línia]: les biblioteques nacionals i els dipòsits digitals nacionals". *BiD: textos universitaris de biblioteconomia i documentació*, desembre 2005, núm. 15. Barcelona: Facultat de Biblioteconomia i Documentació. [Data de consulta: 05/05/2006]. Disponible a: http://www2.ub.edu/bid/consulta_ articulos.php?fichero= 15lluec1. htm>.

[15] *Dipò*sit és la paraula catalana normalitzada que designa el *repository* anglès

[16] LAMARCA, DOLORS. "La Biblioteca de Catalunya en el sistema bibliotecari de Catalunya [En línia]". *BiD: textos universitaris de biblioteconomia i documentació*, juny de 2004, núm. 12. Barcelona: Facultat de Biblioteconomia i Documentació. [Data de consulta: 05/05/2006]. Disponible a: <http://www2.ub.es/ bid/consulta_articulos.php?fichero=12lamarc.htm>.

3.1.2. Models de dipòsits nacionals

Els projectes de dipòsits nacionals han anat dibuixant dos tipologies de models en funció de com es realitza la captura dels recursos digitals.

Per una banda va sorgir el *model integral* o *exhaustiu*, protagonitzat per països com Suècia, Noruega, Finlàndia, Islàndia, Àustria, entre d'altres, el qual es caracteritza per una captura automatitzada dels recursos. Cal matisar, que tot i tractar-se d'una captura automàtica, aquesta està sistematitzada sota uns criteris, habitualment caracteritzats per una zona geogràfica, idioma o domini, però on el productor del recurs té un paper passiu.

Per altra banda tenim el *model selectiu*, que es caracteritza per seguir una política de captura de recursos de forma selectiva i manual (molt allunyada de perseguir un fons exhaustiu), seguint criteris temàtics, geogràfics o d'altres d'interès nacional. La captura es sistematitza amb acords amb les institucions productores facilitant la integració i rebuda dels recursos. Aquest model està representat pels projectes duts a terme per Austràlia, Canadà, Japó, Regne Unit, entre d'altres.

Els límits entre els dos models de dipòsits digitals s'han anat diluint al llarg dels anys per donar pas a un tercer model, l'híbrid, el qual s'està imposant en els països que actualment estan començant aquest tipus de projectes. El model híbrid es caracteritza per una captura periòdica i massiva d'una part de la web (basada amb els criteris anteriorment esmentats), complementada amb acords amb part dels organismes productors. Aquest model reflecteix la doble voluntat de garantir la preservació exhaustiva de la producció bibliogràfica digital nacional i la creació de col·leccions temàtiques relacionades amb esdeveniments d'interès públic.

Altres estudis realitzats sobre els dipòsits nacionals de recursos digitals han observat diferents matisos en relació als models de captura. Aquests matisos donen lloc a una classificació més complexa en funció de la naturalesa dels recursos, ja siguin aquests estàtics o dinàmics[17], però en tot cas, aquesta diferenciació no és excloent amb els casos dels models proposats anteriorment.

Ciro Llueca, coordinador del projecte PADICAT, ha publicat un estudi sobre els models que segueixen els projectes que s'estan duent a terme actualment en aquest àmbit[18].

---

[17] Cordón, José Antonio. "El depósito legal y los recursos digitales en línea [En línia]". A: *Las bibliotecas nacionales del siglo XXI*. Valencia: Biblioteca Valenciana, 2005. [Data de consulta: 06/05/2006]. Disponible a: <http://bv.gva.es/documentos/Ponencias/Cordon.pdf>.

[18] Llueca, Ciro. *Webs sempre accessibles*. Op. cit.

Destaquem l'anàlisi que es fa dels avantatges i inconvenients de cada model, del qual se n'extreu una convergència cada vegada més clara cap un model híbrid (*Annex 2*).

3.1.3. Organitzacions i projectes suprainstitucionals

El desenvolupament dels projectes que s'estan portant a terme en els diferents països està recolzat per l'activitat de diverses organitzacions de caire internacional que tenen la missió d'oferir un marc de cooperació i intercanvi d'experiències. Aquestes organitzacions es caracteritzen per ser consorcis de biblioteques que lideren projectes similars o institucions que han estat pioneres en l'àmbit de la preservació digital:

- *International Internet Preservation Consortium*[19] (IIPC): consorci format per 11 biblioteques nacionals[20] i l'organització Internet Archive, el qual està liderat per la Biblioteca Nacional de França. L'IIPC 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. Dins el seu àmbit d'activitat hi ha la coordinació de diferents grups de treball que investiguen en matèria d'accés i captura de recursos web i la gestió dels seus continguts. En aquest terreny IIPC posa a l'abast programaris i estudis als organismes que es dediquen a desenvolupar projectes de captura del web. Actualment el PADICAT és membre de l'IIPC.

- *Nordic Web Archive*[21] (NWA): fòrum integrat per les biblioteques nacionals escandinaves (Dinamarca, Finlàndia, Islàndia, Noruega i Suècia) amb la finalitat d'intercanviar d'experiències en la coordinació de projectes de captura de recursos del web. Com altres organitzacions, NWA ha desenvolupat un programari en codi obert destinat a la recopilació de recursos, basat en els protocols HTTP i XML per tal de facilitar la comunicació entre sistemes. Aquest

[19] INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. *Netpreserve.org* [En línia]. França: National Library of France, setembre de 2006. [Data de consulta 12/10/2006]. Disponible a: <http://www.netpreserve.org/about/index.php>.

[20] Les onze biblioteques nacionals que integren la IIPC són: Bibliothèque Nationale de France (http://www.bnf.fr) (coordinador), Biblioteca Nazionale Centrale di Firenze (http://www.bncf.firenze.sbn.it), Det Kongelige Bibliotek (http://www.kb.dk), Helsingin yliopiston kirjasto-Suomen Kansalliskirjasto (http://www.lib.helsinki.fi), Kungliga biblioteket Sveriges nationalbibliotek (http://www.kb.se), Landsbokasafn Islands- Haskolabokasafn (http://www.bok.hi.is), Library and Archives Canada (http://www.collectionscanada.ca), Nasjonalbiblioteket (http://www.nb.no), National Library of Australia (http://www.nla.gov.au), The British Library (http://www.bl.uk) i The Library of Congress (http://www.loc.gov).

[21] Nordic Web Archive. *NWA* [En línia]. Noruega: 2006. [Data de consulta: 12/10/2006]. Disponible a: <http://nwa.nb.no/>.

programari, i els seus manuals corresponents, està a la disposició de qualsevol organització interessada en implementar-lo.

- *Internet Archive*[22] (IA): organització sense ànim de lucre nascuda el 1996 a San Francisco amb la voluntat d'esdevenir la *biblioteca d'Internet* a partir de la recopilació dels recursos del web, i posar-los a l'abast dels investigadors, historiadors, personal acadèmic i públic en general. Actualment es considera l'arxiu web més gran del planeta, incloent una gran varietat de formats (text, àudio, imatge en moviment, programari, etc.) accessibles en línia. Rep el suport de diversos organismes, com la Library of Congress, els US National Archives, els UK National Archives, entre d'altres. Internet Archive és un referent clau per la seva experiència i per la producció de programes de captura com Heritrix o de visualització com Wayback Machine.

A part de les organitzacions, que s'han descrit en els anteriors paràgrafs, orientades a donar suport al desenvolupament dels projectes de dipòsits web, cal tenir en compte altres recursos que poden oferir informació molt valuosa de les innovacions i experiències que es van succeint en aquest terreny:

- *PADI*[23] (Preserving Access to Digital Information): es tracta d'un portal realitzat per la Biblioteca Nacional d'Austràlia que ofereix una àmplia informació sobre tots els projectes de preservació del web que s'estan duent a terme en l'actualitat. En aquest portal també podrem trobar informació molt diversa relacionada amb tots els aspectes que cal tenir en compte a l'hora de planificar un projecte de dipòsit de webs, com formats, propietat intel·lectual, polítiques de preservació, programari, etc. A banda d'aquest portal, la Biblioteca Nacional d'Austràlia ha disposat en línia bona part de les presentacions del congrés que va organitzar el novembre de 2004, *Archiving web resources*[24], en el qual s'exposa l'estat de la qüestió de la preservació dels recursos del web.

---

[22] INTERNET ARCHIVE. *Internet Archive* [En línia]. San Francisco: 2006. [Data de consulta: 12/10/2006]. Disponible a: <http://www.archive.org/about/about.php>.

[23] BIBLIOTECA NACIONAL D'AUSTRÀLIA. *Preserving Access to Digital Information* [En línia]. Canberra: Biblioteca Nacional d'Austràlia, 2006. [Data de consulta: 13/10/2006]. Disponible a: <http://www.nla.gov.au/ padi/index.html>.

[24] BIBLIOTECA NACIONAL D'AUSTRÀLIA. *Arching web resources* [En línia]. Canberra: Biblioteca Nacional d'Austràlia, novembre 2004. [Data de consulta 13/10/2006]. Disponible a: <http://www.nla.gov.au/webarchiving/>.

- *IWAW*[25] (International Web Archiving Workshop): es tracta d'unes jornades que des de l'any 2001 es realitzen anualment i que congreguen els professionals que coordinen projectes de preservació del web. Les actes es poden consultar en línia, on podem trobar la informació més actualitzada de l'estat dels projectes i les últimes innovacions. Aquest, com la resta de recursos, representa una font molt important d'informació per observar l'evolució de les tendències i sobretot per conèixer noves eines tecnològiques que faciliten l'execució de les diferents tasques de captura, tractament i difusió de l'arxiu web.

3.2. Anàlisi de projectes de dipòsits de recursos web

La planificació d'un dipòsit de webs exigeix un ampli anàlisi i revisió dels projectes que s'han estat duent a terme fins el moment. La curta existència d'aquest tipus de dipòsits i la contínua evolució de l'entorn web converteix les experiències anteriors en pilars bàsics per la planificació d'un nou projecte. En aquest àmbit ser pioner representa un gran volum d'esforços creant i definint les estratègies per assolir els objectius, per aquest motiu, és vital estudiar tot el que s'ha realitzat fins el moment per evitar fracassos innecessaris. Lògicament els projectes més avançats, i que per tant són el nostre punt d'atenció, són aquells que es van iniciar amb més anterioritat, i que han passat per les fases necessàries per consolidar els seus sistemes.

Cadascun dels projectes defineix les seves característiques en funció dels seus objectius, entorn (lingüístics i geogràfics) i dels recursos dels que disposa, per aquesta raó, tot i tenir una finalitat similar, presenten diferències entre ells. Avui en dia podem comptar amb nombrosos projectes (més o menys consolidats) que tenen com a objectiu crear dipòsits de recursos electrònics en línia (*Annex 3*).

En aquest apartat s'han analitzat sis projectes que realitzen unes funcions similars a les del projecte que es presenta, és a dir, la selecció de recursos en línia amb la finalitat de crear col·leccions temàtiques definides. S'han de remarcar una sèrie de limitacions trobades al llarg d'aquest anàlisi fruit de la quantitat de documentació sobre els diferents projectes, o bé la impossibilitat d'accedir als dipòsits, per no estar en obert, com ha estat en el cas de Netarkivet. Per aquesta raó es planteja un anàlisi general de cada projecte, en funció de la informació disponible, destacant els aspectes favorables i desfavorables de cadascun, posant l'èmfasi en els següents indicadors:

[25] IWAW. *6th International Web Archiving Workshop* [En línia]. 2006. [Data de consulta: 12/12/2006]. Disponible a: <http://www.iwaw.net/06/>.

- Programari
- Limitacions de la infraestructura tecnològica
- Política de captura de recursos digitals
- Política de tractament i descripció dels recursos capturats
- Política d'accessibilitat dels recursos
- Gestió de les col·leccions temàtiques
- Funcionalitats de la interfície de consulta

3.2.1. PANDORA

PANDORA és el projecte liderat per la Biblioteca Nacional d'Austràlia i és el paradigma dels dipòsits nacionals que segueixen el model selectiu. Es tracta d'un dels projectes més veterans, s'inicia el 1996 amb la missió de garantir l'accés permanent i la preservació de la producció digital d'aquest país. El seu abast es centra en la selecció de publicacions en línia i webs sobre Austràlia, d'autors australians o sobre un tema relacionat amb Austràlia. El gran volum d'informació que representa l'abast de PANDORA i la complexitat de construir un arxiu d'aquest tipus, va fer que la Biblioteca Nacional d'Austràlia convidés a altres biblioteques i agències a participar en aquest projecte[26]. El dipòsit de PANDORA està per sobre dels 34 milions de fitxers (desembre de 2006) i té un creixement mensual per sobre del 30 Gb.

Com la majoria de països del món, Austràlia té una llei de dipòsit legal (la vigent és del 1968) que no contempla els recursos digitals, per tant la Biblioteca Nacional amb els seus socis de projecte, mitjançant un comitè científic, van crear una política selectiva de recursos, la qual es fonamenta en la firma d'acords amb els autors dels documents susceptibles de ser capturats. Aquesta política es materialitza en forma de guia on s'especifica la selecció de recursos per part de cada biblioteca integrant del projecte PANDORA.

Per donar suport al creixent nombre de dades el projecte PANDORA va desenvolupar l'any 2001 el programari propi PANDAS. Aquest programa facilita la ingestió dels

[26] Les biblioteques participants al projecte PANDORA, juntament amb la Biblioteca Nacional d'Austràlia són: Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian War Memorial, National Film and Sound Archive, National Library of Australia, Northern Territory Librarys, State Library of New South Wales, State Library of Queensland, State Library of South Australia, State Library of Victoria i State Library of Western Australia

recursos i el seu posterior tractament, indexació i catalogació. PANDAS és un programa que evoluciona en funció de les necessitats de PANDORA, així doncs, l'any 2002 i 2006 es van llançar la segona i tercera versió respectivament, millorant el rendiment de les anteriors.

Pel que fa a les tasques de tractament dels recursos capturats, el fet que es segueixi el model selectiu possibilita poder fer una catalogació detallada, permetent la introducció d'aquests registres en altres catàlegs. La catalogació es fa en format MARC, i aquesta és realitzada en línia pels diferents membres del projecte. PANDORA ha creat un manual de catalogació de recursos digitals per tal d'aconseguir la coherència i qualitat necessària en la descripció.

PANDORA està en accés obert, només un 2% dels seus recursos tenen l'accés limitat per raons comercials. Hi ha diversos punts d'accés per consultar l'arxiu de PANDORA. Per una banda es poden fer consultes des del catàleg de la Biblioteca Nacional d'Austràlia. Per altra banda hi ha la pròpia web del projecte que disposa d'uns índex alfabètics i per matèries dels recursos, i d'un apartat de cerca simple i avançada a text complet. En el cas de la cerca avançada, aquesta permet l'ús d'operadors booleans i concretar la consulta mitjançant criteris de matèria, cronològics i de domini (*Annex 4*).

Del projecte PANDORA cal destacar, per la relació que té amb aquest projecte, l'apartat de col·leccions temàtiques. Quan s'entra en les diferents categories temàtiques es poden observar en un espai destacat les col·leccions de recursos relacionats amb esdeveniments específics. Així doncs, si s'entra per exemple en la categoria d'esports es poden trobar fins a 11 col·leccions temàtiques relacionades amb diferents fets dins la societat australiana. En funció de l'amplitud de la temàtica aquestes col·leccions poden tenir subcategories (*Annex 5*).

Cada col·lecció té una estructura diferent en funció del seu contingut, per tant el nombre de subcategories i recursos serà variable. Les col·leccions tenen una breu explicació del seu contingut general i s'indica quin dels membres del projecte l'ha capturat i en quines dates es va fer.

Els webs capturats es poden navegar en la seva totalitat, però el fet que es tracti d'una col·lecció de recursos seleccionats fa que part de l'estructura dels enllaços externs es perdi. En aquests casos el programa està preparat per avisar a l'usuari d'aquest inconvenient, i el remet a la pàgina vigent del recurs enllaçat.

Valoració de l'observació

Aspectes favorables

- ↑ El fet de tenir un programa propi de gestió de recursos els permet una gran llibertat a l'hora de configurar les seves necessitats i actuar sense limitacions preestablertes
- ↑ Els acords amb els autors dels documents digitals proporcionen una gran solidesa al projecte, ja que permet superar la manca de la llei de dipòsit legal, alhora que facilita l'adquisició contínua dels recursos i l'obtenció de dades necessàries per fer una descripció completa del recurs
- ↑ La catalogació detallada dels recursos confereix un gran valor a la tasca de recopilació d'aquests recursos digitals, ja que s'obté un excel·lent treball bibliogràfic i permet poder usar la màxima potencialitat de les eines de cerca
- ↑ L'accés dels recursos a través del catàleg de la Biblioteca ofereix un nou punt d'accés a tota la informació recopilada per PANDORA, i alhora posa els recursos digitals al mateix nivell de la resta de tipologies documentals
- ↑ El fet que PANDORA es pugui consultar en obert, gràcies als acords amb els autors, amplia el panorama de difusió de tota la seva activitat. És molt important que un projecte d'aquesta envergadura pugui arribar al màxim nombre d'usuaris possibles ja que pot resultar una eina de gran utilitat
- ↑ El sistema de cerca és excel·lent i molt complet, ja que permet la recuperació de recursos tant a través de la navegació com a través de la interrogació
- ↑ Hi ha un gran nombre de col·leccions temàtiques sobre esdeveniments, algunes de les quals s'actualitzen anualment, cosa que permet veure l'evolució sobre un determinat fet

Aspectes desfavorables

- ↓ Gran cost econòmic resultant dels recursos humans i el desenvolupament d'un programari propi
- ↓ El model selectiu del projecte fa que l'univers de recursos preservats sigui només una petita mostra de la realitat
- ↓ El fet de seleccionar recursos de forma individual i aïllada provoca que es perdi l'estructura d'enllaços

3.2.2. Netarkivet

Netarkivet s'origina al 1998 amb la proposta de la llei de dipòsit legal a Dinamarca. El projecte és liderat per la Biblioteca Nacional danesa conjuntament amb la State and University Library. Netarkivet segueix el model híbrid, basat en la captura exhaustiva de la web danesa, la selecció de recursos mitjançant acords amb els autors i la recopilació de documents digitals relacionats amb activitats vinculades a la realitat danesa.

La finalitat del projecte és aconseguir una visió conjunta de la web danesa per tal de ser preservada perpètuament. A partir de la recopilació selectiva de recursos es cobreixen determinades webs amb major freqüència[27], donant una imatge ininterrompuda d'una petita mostra dels llocs web particularment importants i dinàmics. Per altra banda, les altres dues estratègies capturen una major quantitat de webs amb una freqüència més amplia. El conjunt d'estratègies de captura de recursos digitals del Netarkivet es poden expressar gràficament amb la següent imatge[28]:

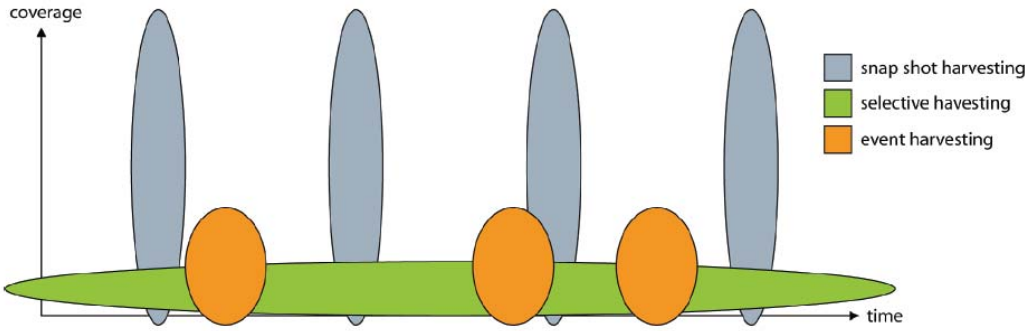


Fig. 8. Estratègies de captura del Netarkivet

Amb l'aprovació de la llei de dipòsit legal, l'1 de juliol de 2005, s'executa la primera captura integral de la web danesa. Durant aquesta primera fase, que va durar de juliol a octubre es van descarregar aproximadament 600.00 dominis, xifra que segons la mateixa institució representa un 60% de la web danesa. Paral·lelament, des de la

[27] Un exemple de l'alta freqüència de captura mitjançant els acords amb els editors és la prova pilot que es va fer amb el diari digital danès *Jyllans-Posten* (JP), una de les principals publicacions en línia del país. El diari en qüestió s'actualitza nombroses vegades al dia, i un dels objectius del projecte és poder captar cadascuna d'aquestes actualitzacions per la importància del seu contingut. L'estratègia per aconseguir totes aquestes captures es basava en la rebuda de la notificació de l'actualització del diari per tal d'executar una captura. CHRISTENSEN-DALSGAARD, BIRTE. "Web Archive Activities in Denmark" [En línia]. *RLG Diginews*. Califòrnia: RLG, Vol. 3 núm. 8., 15 de juny de 2004. [Data de consulta: 28-11-2006]. Disponible a: <http://www.rlg.org/en/page.php?Page_ID=17661#article0>.

[28] ANDERSEN, BJARNE. *The DK-domain: in words and figures* [En línia]. Uhus: State & University Library, 2006. [Data de consulta: 28/11/2006]. Disponible a: <netarchive.dk/publikationer/DFrevy_english.pdf>.

posada en funcionament del projecte fins el 2004, s'havien descarregat 500 Terabytes (Tb) mitjançant la selecció manual de recursos. En el cas de la creació de col·leccions temàtiques, la primera prova pilot es va realitzar en motiu de les eleccions municipals de 2005 durant els mesos d'octubre i novembre.

Al llarg del la fase de disseny del sistema d'informació del Netarkivet es van sotmetre diferents programes a prova. Inicialment es van provar els programes NEDLIB i WGET. Actualment es fan servir els programes HTTrack per la captura selectiva i Heritrix per la captura integral. En el cas d'aquest últim, s'han estat fent una sèrie de desenvolupaments, amb la cooperació de la Nordic National Libraries i l'IIPC, per adaptar-lo a les necessitats del Netarkivet. Les proves pilot de captura van revelar dues limitacions que s'havien de tenir en compte a l'hora d'executar la recopilació integral. Per una banda es va detectar que l'acció de descàrrega de les webs podia saturar els servidors, per aquest motiu es va establir un límit de 5000 pàgines per domini per assegurar la seva integritat[29]. Per altra banda es va descobrir que hi havia determinats webs que tenien incrustats programes[30] que provocaven que el *crawler*[31] entrés en bucle.

Netarkivet no és consultable en línia. El projecte no plantejava un interfície de consulta, ja que aquesta es realitza des del NWA (Nordic Web Archive). Donat que el NWA tampoc té accés en obert, no es poden oferir dades de la interfície de consulta i la visualització dels webs capturats.

Valoració de l'observació

Aspectes favorables

- ↑ Excel·lent estratègia de captura de recursos. El model híbrid, que combina exhaustivitat i selecció, permet crear una imatge fidedigna del web danès, i fer un seguiment dels llocs web que poden resultar més interessants

[29] Aquesta mesura de seguretat per tal de protegir els servidors només afecta a 1% del dominis danesos, ja que la resta estan formats per menys de 5000 pàgines. La majoria de dominis, un 70%, estan composades d'una a deu pàgines.

[30] L'exemple més freqüent d'aquest tipus de programes són els calendaris programats en JavaScript, els quals es caracteritzen per la seva infinitud cronològica basada en la seva naturalesa numèrica, comportant la descàrrega sense límit per part del programa de captura.

[31] *Crawler* és un programa que inspecciona les pàgines del World Wide Web de forma metòdica i automatitzada. Els *Web crawlers* es fan servir per crear una còpia de totes les pàgines web visitades per a ser processades posteriorment per un motor de cerca que indexa les pàgines proporcionant un sistema de cerca ràpid. Els programes *web crawler* comencen visitant una llista de URLs, identifica els hipervíncles d'aquestes pàgines i els afegeix a la llista de URLs a visitar de manera recurrent d'acord amb un conjunt de regles.

- ↑ L'alta freqüència de captura de recursos permet traçar una línia contínua en l'evolució del web, evitant d'aquesta manera les llacunes d'informació dins l'arxiu
- ↑ L'existència d'una llei de dipòsit legal que legitima l'existència del projecte ofereix una plataforma de treball més sòlida i la superació dels obstacles legals en els que es veuen atrapats molts projectes similars
- ↑ La cooperació amb diverses institucions de caràcter supranacional a l'hora de desenvolupar el software ha facilitat la creació d'un sistema d'informació consistent
- ↑ La detecció de certes limitacions a l'hora de capturar els recursos ha permès establir un ritme de treball alt i assegurar la integritat dels servidors

Aspectes desfavorables

- ↓ El fet que Netarkivet no estigui en obert limita molt les seves possibilitats de difusió i utilització per part dels usuaris

3.2.3. UK Web Archive

L'any 2004 neix el projecte UK Web Archive amb la voluntat de preservar i fer accessible de forma permanent la informació digital en línia produïda al Regne Unit, prioritzant els materials de recerca i els llocs web que reflecteixen la cultura britànica[32]. UK Web Archive és un nou exemple de dipòsit basat en la selecció de recursos, i de fet segueix de forma molt fidel, tant pel que fa a la gestió com per l'aparença, el model australià PANDORA.

Per portar a terme aquest projecte l'any 2003 es va crear el UKWAC (UK Web Archiving Consortium), el qual està format per sis institucions[33], amb la finalitat de compartir costos, recursos i l'experiència necessària per assolir els següents objectius:

- Negociar una llicència amb la National Library of Australia per la utilització del programa PANDAS

[32] UKWAC. *UK Web Archiving Consortium* [En línia]: *Archive*.United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/>.

[33] El UK Web Archiving Consortium està format per: la British Library, la qual lidera el projecte, The Nationals Archives, la National Library of Wales, la National Library of Sctoland, el JISC i el Wellcome Trust.

- Creació d'un fòrum de treball en línia, per tal de permetre la discussió sobre la selecció de recursos, la gestió dels drets d'autor i la preservació digital
- Creació d'un marc comú de polítiques de selecció
- Creació d'un protocol comú d'emmagatzematge, tractament i accés als recursos digitals
- Creació d'una interfície en línia que permetí navegar i cercar dins el dipòsit de les webs capturades
- Avaluació de les tasques realitzades i posterior redacció d'un informe amb els resultats obtinguts i les recomanacions per futures accions

El projecte es va posar en funcionament el juny de 2004 amb la selecció i la posterior captura dels recursos digitals. Un any més tard el dipòsit es va disposar en accés obert per a tots els internautes, mentre es continuaven recopilant webs. El juny de 2006 es dóna per acabat el desenvolupament del projecte, amb un resultat de 5.344 webs individuals capturades[34].

Els processos de selecció, captura, emmagatzemament, tractament i preservació dels recursos digitals es realitzen a través del programari PANDAS, el qual possibilita una plataforma de treball descentralitzat consistent. El UKWAC va optar per la utilització del programari PANDAS al tractar-se aquest d'un sistema d'eficàcia comprovada en el marc d'un projecte molt similar al UK Web Archive. Els membres del UKWAC van aprofitar els avantatges d'un sistema completament desenvolupat i flexible, reduint així els costos i temps d'implementació, per introduir-hi certes modificacions que els permetés maximitzar les oportunitats de captura de la web britànica.

Les innovacions que es van aplicar al programa base estaven relacionades amb el protocol de captura de recursos. Per una banda es va voler minimitzar l'efecte advers del *crawler* cap als servidors hostes dels recursos, reduint el nombre de connexions dins un període de temps específic. Per altra banda, es va crear un sistema de seguretat en la ingestió de recursos, ja que al tractar-se d'un sistema distribuït de treball, es podia provocar el col·lapse del dipòsit central. Un altre aspecte clau introduït pel UKWAC va ser l'exclusió del motors de cerca dins el dipòsit, usant la fórmula del robot.txt, per tal d'evitar la confusió de les versions de les webs arxivades amb les

[34] UKWAC. *UK Web Sites Maintained only in the UKWAC Archive* [En línia]: *Report Number 2 September 2006*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://info.webarchive.org.uk/mia_reports/ukwac_mia_list_sept_2006.pdf>.

webs actives. Aquest tipus de polítiques aplicades per la UKWAC asseguren una excel·lent relació amb els propietaris de les recursos digitals i auguren un creixent marc de col·laboració.

Pel que fa als processos de creació de la col·lecció, aquests segueixen el model clàssic de selecció, adquisició, descripció i accés. La captura dels recursos es realitza de forma individual per part de cada membre del consorci[35]. Quan un dels membres decideix entrar un recurs, primer ha de mirar en el mòdul de gestió del PANDAS que aquest no hagi estat entrat o estigui pendent de tractar per un altre membre. En el cas que sigui nou, la institució en qüestió es fa responsable de contactar amb l'autor per aconseguir el permís de captura i seguir els protocols per arxivar el web. Una vegada capturats els recursos s'emmagatzemen al dipòsit central del PANDAS i cada membre s'encarrega de la catalogació dels webs dels que són responsables com una part més de la seva pròpia col·lecció. Aquest sistema ofereix dos avantatges: en primer lloc el dipòsit central permet la navegació i cerca com un sistema autònom; en segon lloc, l'identificador persistent d'URLs usat per localitzar cada web arxivat en el dipòsit central permet ser agregat als catàlegs locals, i així poder ampliar els punts d'accés a les col·leccions.

El UK Web Archive contempla també la possibilitat de dipòsit voluntari de webs a través d'un formulari disponible a la seva pròpia web. En aquest formulari s'adverteix que no totes les webs són incloses immediatament a la col·lecció, ja que aquestes han de ser avaluades per l'agència apropiada del UKWAC.

Pel que fa a la interfície de cerca, tal com s'ha comentat anteriorment, presenta una gran similitud amb el projecte PANDORA, tant per la seva aparença com pel seu funcionament. UK Web Archive permet localitzar recursos de dos maneres: per una banda hi ha un sistema d'interrogació per paraula clau que busca a text complet a totes les pàgines dels documents (*Annex 6*). A diferència de PANDORA, UK Web Archive no ofereix la possibilitat de fer cerques avançades. Com a resultat obtenim un llistat de webs on s'especifica el nom de la pàgina, un breu fragment del text on s'ha localitzat el terme, l'autor i la data de captura. L'altra sistema de localització de recursos funciona a través de la navegació d'una classificació composada per diferents categories i subcategories. En cadascuna d'aquestes categories es mostren les

[35] Cada membre del UKWAC s'especialitza amb una parcel·la temàtica de recursos digitals per tal de facilitar la selecció. Així doncs, per exemple, la Wellcome Library es centra en recopilar webs relacionats amb la medicina, la National Library of Wales webs relacionats amb l'actualitat a Gales i la British Library recull llocs webs relacionats amb fets culturals, històrics o polítics rellevants.

diferents webs que en formen part i les col·leccions temàtiques de recursos relacionats amb esdeveniments. (*Annex 6*)

Quan s'accedeix en el registre d'una web en particular es mostra la informació de si aquesta pertany a alguna col·lecció temàtica, quin dels membres del UKWAC es responsable de la seva captura, un enllaç cap a la web activa i un llistat de les diferents captures efectuades del web. Les captures arxivades permeten una navegació total del lloc, però en molts casos, fruit dels inconvenients del model selectiu, els enllaços externs no estan actius.

En referència a les col·leccions temàtiques s'ha de destacar la seva importància dins el conjunt del projecte, ja que en els primer dos anys s'han creat 17 col·leccions sobre diferents esdeveniments i temes en concret. Aquestes col·leccions tenen una mitjana de 15 webs cadascuna, havent-n'hi algunes amb un nombre molt més elevat, com les que tracten fets com els atacs terroristes a Londres el 7 de juliol de 2005[36] o les eleccions generals 2005 al Regne Unit[37]. La periodicitat del les captures es força alta, donant-se casos de webs que es recopilen gairebé diàriament, com les que formen la col·lecció de la reunió del G8 a Escòcia el 2005[38].

Valoració de l'observació

Aspectes favorables

- Creació d'una plataforma tecnològica de treball distribuït que facilita el treball col·laboratiu per part de tots els membres del consorci
- Adopció d'un programari completament desenvolupat, que redueix els costos i la inversió de temps
- Introducció d'adaptacions al programari base per tal de regular la descàrrega dels servidors dels autors
- Possibilitat dels productors de recursos digitals de realitzar el dipòsit voluntari
- Sistema identificador únic dels recursos per tal de poder-los donar diversos punts d'accés, com els catàlegs dels membres del consorci

[36] UKWAC. *Terrorist attack - London, 7th July 2005* [En línia]: *related Internet Sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col/c8125.html>.

[37] UKWAC. *General Election - UK 2005* [En línia]: *related Internet Sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col/c8100.html>.

[38] UKWAC. *G8 summit 2005* [En línia]: *related Internet sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col/c8150.html>.

- ↑ Dipòsit sense cap tipus de restricció, obert a qualsevol internauta
- ↑ Ampli ventall de col·leccions relacionades amb diferents temes i esdeveniments
- ↑ Alta periodicitat de captura dels recursos digitals, que permet traçar una contínua evolució dels llocs web
- ↑ Interfície amigable amb diferents eines de localització de recursos: navegació a través de les categories i cerca per paraula clau

Aspectes desfavorables

- ↓ El model selectiu trenca la majoria d'enllaços externs i comportant la pèrdua dels context dels recursos digitals
- ↓ Manca d'un apartat de cerca avançada, per tal de poder aprofitar els avantatges que ens ofereix la catalogació dels recursos

3.2.4. European Archive

European Archive és una fundació sense ànim de lucre, amb seu a Amsterdam i París, que treballa per la creació d'un accés universal al coneixement a través d'Internet, a partir de la digitalització de documents, recopilant el web europeu i creant tècniques que assegurin la preservació d'aquest ric llegat. A diferència dels anteriors projectes, aquest es situa en un àmbit supranacional, per tant la seva finalitat és molt més àmplia, ja que és més complexa definir l'abast.

European Archive compte amb la col·laboració d'Internet Archive i XS4ALL[39] per tal de construir el primer arxiu digital global europeu, i per erigir-se com una institució evocada a donar suport a altres institucions que pretenen crear i preservar les seves col·leccions digitals.

El web del European Archive[40] disposa de diferents tipologies de documents digitals, tots accessibles en obert: audiovisuals (films oferts pel govern britànic), sonors (col·lecció de música clàssica oferta per una ràdio neerlandesa) i multimèdia (recopilació de webs a nivell europeu).

[39] XS4ALL (*access for all*), companyia fundada al 1993, és el principal proveïdor de serveis d'Internet als Països Baixos.

[40] EUROPEAN ARCHIVE FOUNDATION. *European Archive* [En línia]. [Amsterdam], European Archive Foundation, 2006. [Data de consulta: 05/02/2007]. Disponible a: <http://www.europarchive.org/index.php>.

El dipòsit que allotja la col·lecció digital, situat a Amsterdam, té actualment una capacitat de 250 Tb, amb una previsió de multiplicar-se per quatre en els propers anys, per tal de poder amplificar el procés de captura a gran escala. L'última xifra disponible sobre la descàrrega d'objectes digitals de l'arxiu, de desembre de 2005, ofereix una mitjana superior als 350 Megabytes per segon. Per tal de facilitar el treball de gestió i poder donar resposta a les nombroses entrades s'ha dissenyat una infraestructura distribuïda configurada per més 200 nodes.

Actualment es pot accedir a quatre col·leccions temàtiques de recursos web referents a diferents àmbits europeus (*Annex 7*):

- *European Constitution Web Archive*: recopilació de 249 webs de partits polítics de diferents països de la Unió Europea, amb captures executades durant els mesos immediatament anteriors i posteriors a les eleccions de la Constitució Europea. Aquesta col·lecció segueix el model selectiu de captura, tot i que per la limitació de recursos l'abast del projecte no cobreix tots els partits implicats en el debat de la Constitució Europea.
- *UKGOV Weekly Web archive*: recull d'11 llocs web governamentals britànics que són capturats amb freqüència setmanal.
- *UKGOV six monthly web archive*: recull de 57 llocs web governaments, diferents als de l'anterior col·lecció, que són capturats dos vegades l'any. Aquesta col·lecció, com l'anterior, segueix el model selectiu de recursos digitals.
- *Italian Domain Snapshot 2006*: el maig de 2006, juntament amb la col·laboració de la Biblioteca Nacional d'Itàlia, es va realitzar una captura integral i exhaustiva del domini web italià.

Les quatre col·leccions es poden consultar en obert mitjançant diferents interfícies. Les tres primeres estructuren els webs a partir de llistats, mentre que la del domini italià s'ha de cercar a través de la introducció d'URLs. Una vegada seleccionat el web en qüestió es mostra un llistat de dates de captura, en una interfície adaptada del programari Wayback Machine d'Internet Archive, que permeten executar el web. European Archive incorpora un sistema de cerca per paraula clau, però només es pot aplicar a les col·leccions de música i vídeo.

La novetat respecte els altres projectes exposats és el grau d'interacció, ja que permet a l'usuari, previ registre, crear una col·lecció pròpia a partir dels recursos disponibles i gestionar-los mitjançant determinades funcionalitats. Actualment aquestes

funcionalitats encara no estan activades, però ja es presenten algunes de les novetats, com la possibilitat de poder agregar etiquetes (*tags*) per descriure els recursos.

Valoració de l'observació

Aspectes favorables

- ↑ L'aliança amb Internet Archive permet a l'European Archive poder adaptar els recursos tecnològics per reduir costos, i absorbir determinats recursos capturats pel primer
- ↑ Treball col·laboratiu amb diverses institucions europees per la creació de col·leccions a nivell general
- ↑ Accés en obert dels recursos capturats per tal de fomentar l'accés universal del patrimoni cultural europeu
- ↑ Ús dels diferents models de captura de recursos en funció dels objectius i característiques de cada col·lecció
- ↑ Interacció amb l'usuari final amb la creació d'un apartat personal que li permet la gestió de determinats recursos en funció dels seus interessos
- ↑ Ús d'etiquetes per descriure els recursos digitals

Aspectes desfavorables

- ↓ El model selectiu trenca la majoria d'enllaços externs, comportant la pèrdua dels context dels recursos digitals
- ↓ Manca d'un sistema de cerca per paraula clau aplicat a la col·lecció de webs

3.2.5. MINERVA

Sota el lideratge de la Library of Congress, l'any 2000 s'inicia el projecte MINERVA[41], amb la finalitat de recopilar i preservar recursos web, per tal d'evitar la seva pèrdua degut a la seva duració efímera[42].

[41] MINERVA és l'acrònim de **M**apping the **IN**ternet **E**lectronic **R**esources **V**irtual **A**rchive.

[42] Diverses institucions, entre elles la Library of Congress, sustenen que la vida mitja d'una pàgina web és de 44 dies. **Vegeu**: LIBRARY OF CONGRESS. *Web Capture & Archiving* [En línia]. Washington: Library of Congress, abril de 2003. [Data de consulta: 06/02/2007]. Disponible a: <http://www.loc.gov/acq/devpol/webarchive.html>.

MINERVA està format per un equip multidisciplinari, provinent de diverses institucions, especialitzat en les diferents tasques de selecció, captura, catalogació i preservació de recursos. La Library of Congress, actuant com a líder, és qui fixa la planificació de les col·leccions i el seu contingut, a més de donar accés a aquestes i assegurar-ne la seva preservació. Internet Archive, amb la seva llarga experiència en l'àmbit dels arxius web, és l'agent que s'encarrega de la captura i emmagatzematge dels recursos digitals, i el desenvolupament i manteniment del programari Wayback machine per poder accedir al dipòsit. WebArchivist.org[43] s'encarrega del desenvolupament d'eines tecnològiques per poder tractar els recursos capturats, sobretot en el camp de l'agregació de metadades i la creació d'interfícies pels agents editors. Finalment, el Pew Internet & American Life Project, a part d'aportar capital per finançar el projecte, s'encarrega de l'anàlisi i la redacció d'informes sobre l'evolució de MINERVA.

A mitjans de l'any 2000 es va llançar el primer prototip de MINERVA que funcionava amb un incipient *crawler*, amb el qual es van capturar 12 webs. Aquestes van ser catalogades i es van fer accessibles a un campus per tal de realitzar una avaluació dels processos. L'any 2001 MINERVA ja havia definit les seves polítiques de captura del web i el funcionament del seu sistema d'informació, el qual s'ha mantingut fins avui i es pot resumir amb el següent gràfic:

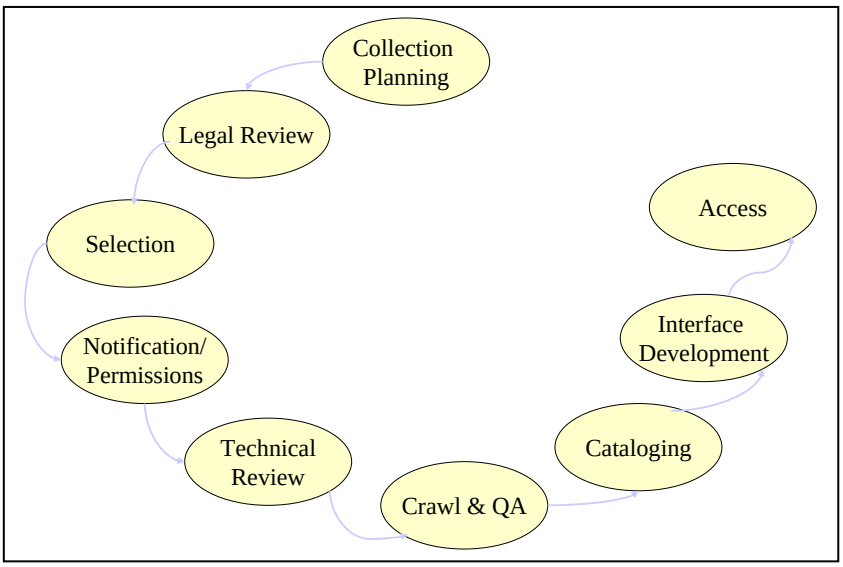


Fig. 9. Processos del projecte Minerva

[43] WebArchivist.org és un grup d'investigació i desenvolupament de software de la University of Washington i del SUNY Institute of Technology que dóna suport a bibliotecaris i arxivers interessats en preservar i analitzar els materials creats i distribuïts pel web.

Tal com s'ha esmentat ens els anteriors paràgrafs, la Library of Congress gestiona la planificació de les col·leccions temàtiques, en la qual es defineix el seu abast i els diferents paràmetres de la selecció: volum i període de la col·lecció (començament i finalització del procés), i la freqüència i profunditat de la captura. Llavors es desenvolupa la política de selecció, en la qual s'identifiquen els webs potencialment interessants en base als criteris temàtics de la col·lecció, que sempre estarà relacionada amb algun esdeveniment en concret.

Donat que la llei de dipòsit legal nord americà no inclou els recursos en línia[44], totes les captures que es realitzen en el si del projecte MINERVA requereixen la notificació a l'usuari i els seu corresponent permís per poder mostrar el recurs en obert i manipular-lo per assegurar la seva preservació.

Una vegada han estat capturats el recursos seleccionats aquests són catalogats en la fase de tractament. MINERVA elabora els registres de metadades mitjançant l'estàndard de descripció MODS (Metadata Object Description Schema)[45] (*Annex 8*). Els camps de descripció dels registres catalogràfic són els següents:

- Títol
- Nom (autoritat controlada)
- Resum
- Data de captura
- Tipologia (per defecte és lloc web)
- Descripció física (format de l'arxiu)
- Identificador (URL)
- Llengua
- Condicions d'accés i drets de gestió
- Matèria (basada en la LCSH)

Pel que fa a l'accés als recursos el projecte MINERVA es caracteritza per desenvolupar interfícies molt elaborades, tant pels punts d'accés que s'ofereixen, com la quantitat d'informació que es recull de cada recurs (metadades). Actualment

[44] U.S. COPYRIGHT OFFICE. *Mandatory deposit (17 U.S.C. section 407)* [En línia]. Washington: U.S. Copyright Office, Juliol de 2006. [Data de consulta: 06/02/2007]. Disponible a: <http://www.copyright.gov/help/faq/mandatory_deposit.html>.

[45] MODS és una *Schema XML* que integrar les dades creades amb MARC 21. L'*schema* està basat en els camps de descripció de MARC 21, però enlloc d'usar el llenguatge numèric de MARC empra una estructura d'etiquetes, les quals permeten ser agrupades.

MINERVA té quatre col·leccions temàtiques en obert[46], i cinc que estan estat de producció[47]. Quan s'accedeix a qualsevol d'aquestes col·leccions el primer que trobem és la secció *About*, en la qual se'ns donen totes les dades referents a la seva creació: descripció, criteris de selecció, metadades usades, detalls tecnològics, informació sobre copyright, preguntes freqüents i els socis que hi han participat. (*Annex 8*)

A nivell d'eines de cerca, en funció de la col·lecció que visitem, podrem només fer servir recursos de navegació (*browse*) o també tindrem la possibilitat consultar a través de paraula clau (*search*). En tot cas, les col·leccions que només ofereixen la possibilitar de navegar els recursos per localitzar el que busquem, donen múltiple punts d'accés al fons de webs, que faciliten substancialment la recerca. Entre els diferents punts d'accés que podem trobar en la interfície d'una col·lecció destaquem:

- Índex alfabètic de tots els webs
- Característiques del productor (govern, personal, religiós, mitjà de comunicació, educació,etc.)
- Localització geogràfica del productor (per països)
- Llengua (fins a un total de 31 llengües)

Quan es selecciona un recurs concret dels disponibles en la llista de navegació, s'obre una nova finestra del navegador on apareixen totes les dades catalogràfiques del lloc web, i la possibilitat d'entrar a l'historial de versions capturades del recurs en qüestió (annex fitxa catalogràfica). Aquest historial es mostra en la típica i ja comuna forma de calendari que ha popularitzat Wayback machine, on es plasmen les dates dels dies que es va capturar el web. L'última acció a realitzar és triar unes d'aquestes dates per poder visualitzar el recurs, el qual també és navegable, menys en el cas dels enllaços externs.

Valoració de l'observació

Aspectes favorables

↑ L'aliança d'institucions que ha creat la Library of Congress per dur a terme el projecte Minerva en garanteix l'èxit, tant per la capacitat d'autofinançar-se com per l'experiència en el terreny de la gestió de recursos digitals dels seus *partners*.

[46] Les col·leccions que estan operatives són les de les eleccions del 2000, els fets de l'11 de setembre, les eleccions del 2002 i la constitució del 107è Congrés.

[47] Les col·leccions que encara estan en fase de producció són les dels Jocs Olímpics d'hivern de 2002, el primer aniversari dels fets de l'11 de setembre, els fets de la guerra d'Irak, la constitució del 108è Congrés i les eleccions de 2004.

- ↑ L'obtenció del permís dels autors dels recursos permet la possibilitat de crear polítiques de preservació que requereixin la manipulació dels arxius.
- ↑ Accés en obert dels recursos capturats per tal de difondre més àmpliament els continguts de les col·leccions
- ↑ Excel·lent política de catalogació dels recursos, la qual permet poder contextualitzar completament els recursos de forma individual, facilitar-ne la localització mitjançant els diferents punts d'accés, i fins i tot agregar-los en un catàleg de Biblioteca, ja que són compatibles amb el format MARC.
- ↑ Navegació de la interfície de cerca molt completa, amb diferents punts d'accés i amb diferents recursos que ajuden a orientar-se
- ↑ Col·leccions riques per la seva quantitat i qualitat de recursos

Aspectes desfavorables

- ↓ El model selectiu trenca la majoria d'enllaços externs, comportant la pèrdua dels context dels recursos digitals
- ↓ Manca d'un sistema de cerca per paraula clau en algunes col·leccions
- ↓ Pel potencial dels integrants de Minerva hi haurien d'haver més col·leccions temàtiques

3.2.6. WebArchiv

WebArchiv[48] va ser iniciat com un projecte de R+D finançat pel Ministeri de Cultura Txec l'any 2000, i que des d'aquell moment va continuar desenvolupant-se en el si de la Biblioteca Nacional de la República Txeca, juntament amb la Biblioteca de Moràvia, responsable dels aspectes tecnològics, i la Universitat de Masaryk, com a col·laborador extern.

La missió del projecte WebArchiv és la creació i implementació d'una solució tecnològica en el camp de la preservació del *web nacional*, sobretot en els casos dels documents que només estan en format digital (*born digital*). Els objectius estan centrats en la creació d'eines i mètodes que permetin la captura, allotjament i preservació dels recursos web, de manera que se n'asseguri l'accés de forma perpètua.

[48] ZABICKA, PETR. *WebArchiv* [En línia]: *Czech Web Archive Revisited*. Brno: Moravian Library in Brno, 2006. [Data de consulta: 25/03/2007]. Disponible a: <www.iwaw.net/06/PDF/iwaw06-zabicka.pdf>.

WebArchiv és un clar exemple del model híbrid de dipòsit d'arxius digitals, ja que combina la captura integral del web nacional i la captura selectiva de recursos, juntament amb la creació de col·leccions temàtiques relacionades amb esdeveniments públics en la societat txeca.

Només una petita part dels recursos capturats són mostrats en obert, ja que la legislació de dipòsit legal de la República Txeca no contempla els documents creats únicament en format digital, motiu pel qual els responsables d'aquest projecte han optat per restringir-ne la seva difusió.

Els processos que es realitzen en el WebArchiv estan descentralitzats en les seus de Praga (Biblioteca Nacional de la República Txeca) i Brno (Biblioteca Nacional de Moràvia).

A Praga es porten a terme les tasques de:

- Gestió del projecte
- Selecció de recursos
- Catalogació de recursos amb metadades seguint el model Dublin Core
- Creació d'acords amb els autors de les recursos

A Brno es porten a terme les tasques de:

- Implementació i manteniment del maquinari del projecte
- Instal·lació, manteniment i desenvolupament del programari del projecte
- Anàlisi i simulació de la captura de recursos
- Captura dels recursos seleccionats
- Indexació de recursos capturats
- Facilitar l'accés dels usuaris als recursos, mantenint la interfície de consulta

Durant els anys 2001, 2002 i 2004 es van realitzar diverses captures integrals del domini txec amb el resultat 95 milions d'arxius recopilats, equivalents a 3,6 Tb. Actualment el projecte ja ha capturat més de 134 milions d'arxius i ja supera el volum dels 5 Tb.

Pel que fa al programari, el projecte va començar implementant l'aplicació Nedlib, que funcionava juntament amb les Nordic Metadata project tools. Actualment WebArchiv funciona amb el set d'aplicacions de l'IIPC:

- Heritrix: programa de captura
- Nutch Wax: indexació dels recursos i motor de cerca
- Web based Dublin Core creator;: programa d'agregació de metadades

- Wera: programa d'accés als recursos públics
- Wayback machine: interfície d'accés a la totalitat dels recursos del projecte

A nivell de maquinari, WebArchiv fa servir tres servidors HP Proliant i té un dipòsit amb 5,8 Tb. Està previst que a finals de 2007 es transfereixi tot el fons del dipòsit a la Biblioteca Nacional Txeca, on disposarà de més de 25 Tb d'allotjament.

Des del 2002 WebArchiv ha produït 6 col·leccions temàtiques, de les quals només una part del recursos que les formen es poden accedir en obert[49]. Per accedir als recursos de les col·leccions WebArchiv posa a disposició de l'usuari dos punts d'accés: la cerca per paraula clau i la localització a través d'URL (*Annex 9*). Donat que la majoria de recursos no estan en obert, en la interfície també es pot trobar un camp de consulta dirigit a Internet Archive, on segurament es podran trobar alguns dels webs capturats per WebArchiv però que no es poden trobar en línia.

Valoració de l'observació

Aspectes favorables

↑ Creació d'una plataforma tecnològica de treball distribuït destinada a poder treballar en col·laboració a altres institucions.

↑ Adopció d'un programari catalogat com estàndard per configurar el sistema d'informació del projecte

↑ Col·leccions temàtiques riques pel nombre de recursos

↑ Enginyosa alternativa per mostrar els recursos que no estan en obert a través del projecte Internet Archive

↑ Interfície amigable amb una gran quantitat d'informació sobre la missió i el funcionament del projecte

Aspectes desfavorables

↓ Gran part del fons no està en obert

↓ Manca d'un apartat de cerca avançada, per tal de poder aprofitar els avantatges que ens ofereix la catalogació dels recursos

[49] Les col·leccions temàtiques produïdes pel WebArchiv són: Inundacions a l'Europa central (2002); Adquisició del incunable Chronique of Dalimil (2005); Monogràfic de la regió dels Highlands (2005); Eleccions estatals (2006); Construcció del nou recinte de la Biblioteca Nacional Txeca (2007); i Candidatura de Praga als jocs Olímpics de 2016 (2007). **Vegeu**: WEBARCHIV. *Thematic collections* [En línia]. Praga: Biblioteca Nacional de la República Txeca, 2007. [Data de consulta: 25/03/2007]. Disponible a: <http://en.webarchiv.cz/thematic_collections#vysocina>.

3.3. Anàlisi del projecte PADICAT

L'existència i desenvolupament del projecte TematiCAT s'origina en l'objectiu de crear un sistema de captura selectiva de recursos relacionats amb esdeveniments temàtics dins el marc del projecte PADICAT. Per poder emprendre aquest projecte es requereix un profund anàlisi del projecte PADICAT per tal de realitzar un desenvolupament coherent amb la plataforma que es relaciona. En aquest apartat es presentarà una radiografia dels objectius, funcionalitats, sistemes d'informació i recursos disponibles en el PADICAT, extret de la *Memòria del Plantejament* publicada el desembre de 2005.

Missió

Dissenyar i crear un sistema que permeti a la Biblioteca de Catalunya compilar, processar i donar accés permanent a part de la producció digital catalana. Aquest projecte pren la forma d'un dipòsit de recursos digitals definit pel desenvolupament de tres objectius estratègics:

- Compilació massiva de recursos publicats en obert a Internet
- Impulsió d'un dipòsit sistemàtic de la producció web dels agents implicats a Catalunya
- Promoció de línies de recerca per mitjà de la integració dels recursos digitals de determinats esdeveniments de la vida pública catalana

Les perspectives de la planificació del projecte preveuen que al 2009 ha de permetre a la Biblioteca de Catalunya comptar amb un escenari òptim, pioner a Espanya i de referència a Europa, que funcioni a ple rendiment, amb uns indicadors quantitatius de 100.000 pàgines web capturades en diverses edicions, que possiblement signifiquin uns 50 milions d'arxius i 30 Tb de volum. Paral·lelament, es preveu tancar acords de cooperació amb unes 300 institucions de tot tipus, així com permetre l'accés en obert, en línia, a bona part de la col·lecció[50].

L'assoliment dels objectius marcats pel PADICAT reverteixen uns beneficis que en general engloben tota la societat, i que es poden desglossar de la següent forma[51]:

---

[50] BIBLIOTECA NACIONAL DE CATALUNYA. *PADICAT*: *Patrimoni Digital de Catalunya* [En línia]. Barcelona: Biblioteca Nacional de Catalunya, 2006. [Data de consulta: 24/10/2006]. Disponible a: <http://www.padicat.cat/quees.php>.

[51] Ibíd.

- Per a la ciutadania, accés obert i permanent als recursos que són fruit del coneixement i l'expressió dels creadors del segle XXI, ja siguin de caràcter cultural, educatiu, científic o administratiu, o comprenguin informació tècnica, jurídica, mèdica o d'un altre tipus.

- Per a les institucions, empreses, administracions i particulars que produeixen pàgines web a Catalunya, preservació de la pròpia producció i garantia d'accés, amb els condicionants que la llei regeix, als continguts i dissenys que, altrament, desapareixerien.

- Per al sistema bibliotecari, possibilitats infinites de cooperació amb la resta de biblioteques, arxius i museus de Catalunya; impuls i lideratge en la confecció del patrimoni digital d'Espanya.

3.3.1. Anàlisi de l'abast

L'abast del projecte es centre en la recopilació, processament i difusió d'un conjunt de recursos molt heterogenis, tant pel que fa temàtica com pel que fa a la seva naturalesa (tecnològica), en funció de la seva pertinença al Patrimoni Digital Català.

S'entén com a norma de base que *Patrimoni Digital* és aquella informació publicada a Internet, en obert o no, independentment del format en què es presenta. Per altra banda, s'entén *de Catalunya* en base a la definició de la *producció bibliogràfica catalana* que fa la Biblioteca Nacional de Catalunya, és a dir, tot allò produït a Catalunya o que tracti sobre Catalunya. En la *Memòria del plantejament del projecte PADICAT*[52] s'empra el concepte *comunitat web de Catalunya* per definir aquest àmbit, polític, geogràfic i social.

Anàlisi de l'abast temàtic

Tenint en compte que la font d'informació del PADICAT és Internet, i que aquesta ha estat concebuda per diluir les fronteres i fer accessible la informació de forma universal, s'ha intentat identificar mòduls d'interès de grups concrets o, com s'ha definit anteriorment, *comunitats d'usuaris web*. En funció de l'anàlisi de l'abast d'altres projectes similars, PADICAT centra el seu abast temàtic en el grup de documents que contenen informació relativa a Catalunya o d'interès majoritari de la societat

[52] BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT*. Op. cit.

catalana[53]. En aquest sentit, l'abast temàtic del Patrimoni Digital de Catalunya segueix la següent estratègia:

- Webs sota el domini .cat[54]
- Webs ubicades a servidors de Catalunya
- Webs sota dominis geogràfics (.es, .com, .net, .org, etc.) en llengua catalana
- Webs que no compleixin els requisits anteriors, però relacionades temàticament amb Catalunya

Anàlisi de l'abast tecnològic

S'ha de prestar especial atenció a l'abast tecnològic del projecte, ja que com la mateixa xarxa (Internet), els documents que es recopilen tenen un concepte de finitud molt variable. També és important tenir en compte el dinamisme i evolució dels objectes digitals, ja que dia a dia apareixen innovacions que condicionen aquest abast tecnològic.

En el marc del projecte es presenta una definició de *seu web*[55] que serveix per limitar i fonamentar una base de les característiques que han de complir els recursos que han de formar part del dipòsit del PADICAT:

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

---

[53] *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 (...).* **Llei de Biblioteques de Catalunya, de 24 d'abril de 1981**, *Diari Oficial de la Generalitat de Catalunya*, núm. 123 (29 abr. 1981). *Nota*: Disponible en línia a: Wikisource. *Llei de Biblioteques de Catalunya, de 24 d'abril de 1981* [En línia]. Desembre 2005. [Data de consulta: 10/06/2006]. Disponible a: <http://ca.wikisource.org/wiki/Llei_de_biblioteques_de_Catalunya_1981>.

[54] El domini .cat va ser aprovat per la ICANN (Internet Corporation for Assigned Names and Numbers) el 16 de setembre de 2006.

[55] *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.* **Vegeu**: Pareja, Víctor Manuel; Ortega, José Luís; 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.

Tot i la definició d'un abast tecnològic general, s'han d'observar alguns aspectes que no queden resolts fruit de les limitacions del propi sistema de captura automàtica de recursos. Aquestes casuístiques, les quals s'han de tractar amb un anàlisi més específic dins l'àmbit de l'abast tecnològic, venen donades per la complexitat d'accés als documents, ja sigui perquè contenen aplicacions JavaScript (menús de navegació, informació dinàmica, aplicacions de veu, etc.), o bé perquè són recursos restringits per mitjà de contrasenyes o controls IPs.

Anàlisi dels requeriments tecnològics

El sistema d'informació del PADICAT treballa en base del sistema operatiu Linux, en el qual s'ha implementat el programa servidor web Apache 2 i el programa servidor d'aplicacions Tomcat 5. Sobre aquesta plataforma s'ha implementat els diferents programes que formen l'arquitectura el sistema d'informació del PADICAT:

- Heritrix: programa de captura de recursos
- BAT: programa d'administració i manipulació de recursos en format comprimit (ARC)
- Nutch Wax: programa que efectua les funcions d'indexació dels recursos i que actua com a motor de cerca en l'entorn del dipòsit.
- Wera: programa que genera una interfície de cerca per l'usuari final i que permet visualitzar els recursos resultants

Entre els mesos de desembre de 2005 i gener de 2006, recolzant la fase de disseny i d'implementació de l'arquitectura del sistema d'informació, el projecte PADICAT va comptar amb l'assessorament d'un analista informàtic de la consultora en tecnologia de la informació AUSEBA.

Pel altra banda, el maquinari que suportarà el sistema consisteix en:

- 2 servidors SUN Fire V490 amb un 1 Gb de memòria
- 1 dipòsit de 10 Tb d'espai per emmagatzemament de dades, que es preveu que pugui créixer anualment de 10 Tb addicionals.
- 1 llibreria de cintes LTO de 10 Tb amb un creixement anual de 10 Tb
- 1 servidor estàndard per la plataforma web

Anàlisi dels recursos humans

Respecte als requeriments de recursos humans en el projecte PADICAT, en la *Memòria del plantejament* s'ofereix un estimació dels agents que participaran en el desenvolupament, ja que es contempla des de la fase de disseny a la de producció, tenint en consideració que en funció dels resultats obtinguts, la necessitat de recursos pot ser major o menor. També cal destacar que en la descripció d'aquests requeriments es compte amb l'expertesa i suport del personal de la Biblioteca de Catalunya, així com la dels col·laboradors del projecte (com per exemple els professionals del CESCA). Al llarg de les diferents fases del projecte es preveu la participació dels següents perfils:

- Cap de projecte
- Analista especialitzat, dedicat preferentment a la recollida de recursos
- Analista especialitzat, dedicat preferentment a la gestió i organització dels recursos
- Dissenyador d'interfície web
- 2 Bibliotecaris de suport
- 2 Bibliotecaris especialistes en metadades
- Administratiu

3.3.2. Anàlisi del sistema d'informació

El sistema d'informació del PADICAT té una estructura i funcionament molt similar a un cicle documental clàssic d'un servei d'informació o biblioteca. A grans trets el sistema d'informació està composat per uns mòduls, els quals es tractaran més detalladament en els següents apartats, que executen els processos de captura, tractament i accés permanent dels recursos.

El següent gràfic[56] mostra amb més detall l'arquitectura del sistema, i les diferents accions que s'efectuen en cadascun dels processos.

[56] BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT*. Op. cit.

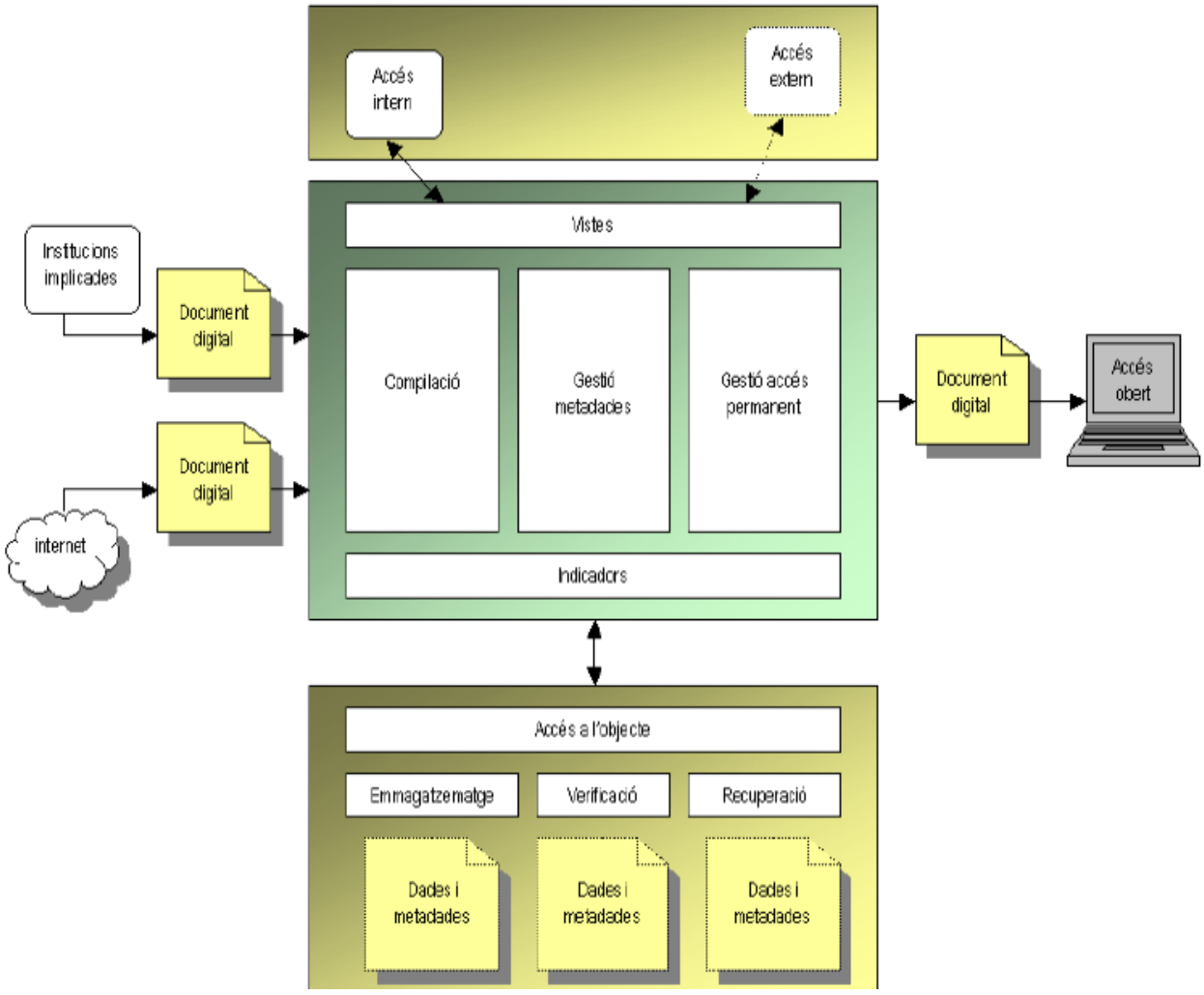


Fig. 10. Esquema dels processos del PADICAT

El sistema està composat per diversos programes de codi obert, on cadascun dels quals realitza un dels processos esmentats, i treballen associats a una base de dades MySQL on s'agrupen els registres relatius als recursos web que formen la col·lecció.

El gràfic[57] que es mostra a continuació permet analitzar les parts components del sistema d'informació i la seva interacció, juntament amb els agents que fan possible el seu funcionament:

[57] Gràfic elaborat per Leandro Stasi, enginyer informàtic del projecte PADICAT. **Vegeu**: BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT*. Op. cit.

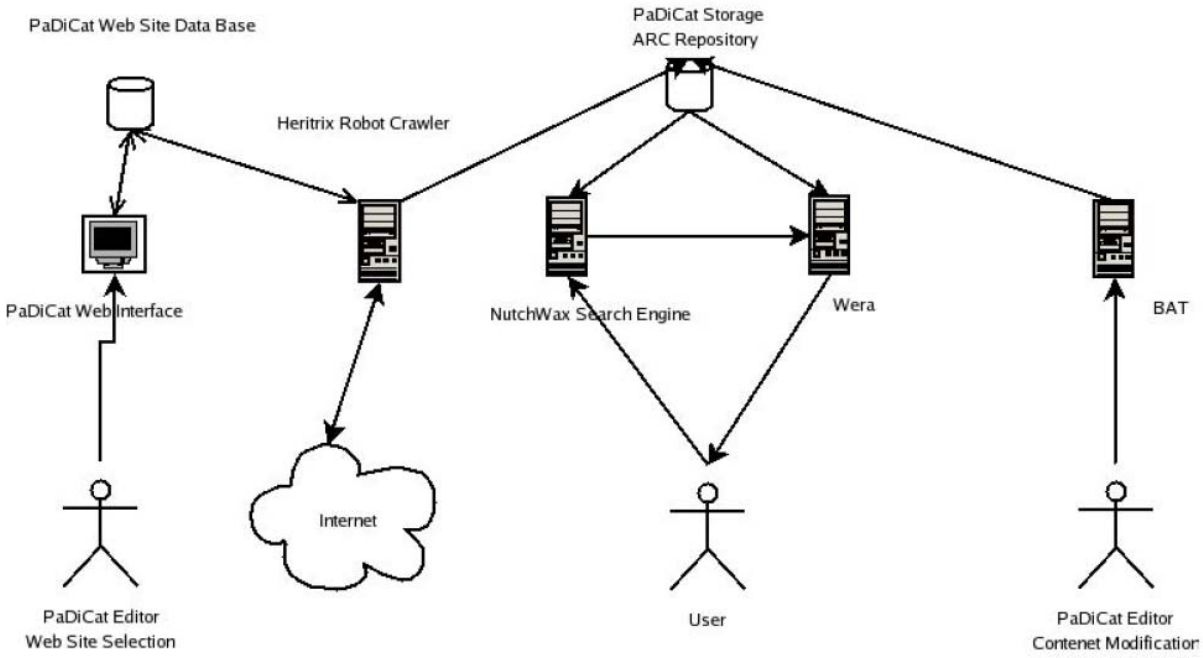


Fig. 11. Estructura dels components del sistema d'informació del PADICAT

- **Editor** [PaDiCat Editor]: aquest agent s'encarrega de la selecció dels recursos i de la seva revisió o descripció. Tot i que en el gràfic no apareix, l'editor o un agent similar, és qui s'encarrega de donar els paràmetres de funcionament al programa o de vetllar per la correcte comunicació entre els components.
- **Usuari extern** [User]**:** aquets és l'agent final que, a través de la interfície de consulta, executa el procés de cerca per tal d'accedir als recursos de la col·lecció allotjada al dipòsit.
- **Interfície de l'editor** [PaDiCat Web Interface]**:** aquesta facilita l'accés i la organització de la col·lecció a l'editor.
- **Interfície de l'usuari** [Wera]: aquest és un programa lliure desenvolupat per la Biblioteca Nacional de Noruega dissenyat especialment per a la visualització de recursos en aquesta tipologia de dipòsits. Wera permet la cerca i recuperació de recursos i el seu posterior accés.
- **Internet:** font d'informació principal d'on s'extreuen els recursos
- **Dipòsit** [PaDiCat Storage ARC Repository]: unitat on es troben allotjats els recursos capturats en un format comprimit.

- **Base de dades** [PaDiCat Web Site Data Base]: inclou els registres dels recursos que formen la col·lecció.
- **Robot** [Heritrix Robot Crawler]: programari de codi obert encarregat de capturar els recursos d'Internet en funció d'uns paràmetres donats.
- **Indexador** [Nutch Wax Search Engine]: programari de codi obert encarregat de la doble funció d'indexar la col·lecció i fer cerques en ella.
- **Sistema d'administració** [BAT]: programari lliure que permet realitzar modificacions als arxius comprimits allotjats en el dipòsit.

3.3.2.1. Captura dels recursos

La captura dels recursos ha estat l'objectiu prioritari de la primera fase del projecte PADICAT, ja que d'aquest procés en depenen accions posteriors i l'avaluació del sistema d'informació. Aquest procés treballa conjuntament amb una base de dades (PaDiCat Web Site Data Base) on s'introdueixen els registres que són recopilats, els quals són revisats i tractats posteriorment per passar a formar part del dipòsit.

La compilació massiva de recursos s'alimenta de quatre vies d'adquisició fonamentades en diferents mètodes:

- Captura automatitzada
- Captura manual
- Creació del dipòsit voluntari sistemàtic de la producció web dels agents implicats a Catalunya
- Integració dels recursos digitals de determinats esdeveniments de la vida pública catalana

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

Fig. 12. Estratègies de captura de recursos del PADICAT

Captura automatitzada

La captura automatitzada es realitza per mitjà d'Heritrix, que és un programa de codi obert desenvolupat en les seves darreres versions per l'Internet Archive i les Nordic National Libraries. Aquest programa ha estat dissenyat per navegar a la xarxa a partir d'unes coordenades prèvies (URLs). D'acord amb els paràmetres que dóna l'administrador del sistema, Heritrix captura els recursos que compleixen els criteris predefinits, oferint la possibilitat de descarregar-los directament en format comprimit. El projecte PADICAT treballa en tres vies de desenvolupament de paràmetres que permeten la captura automàtica dels recursos objectius:

- Webs sota domini .cat: aquest domini va dirigit a la comunitat lingüística i cultural catalana a Internet[58], i per tant, representa una excel·lent oportunitat d'aconseguir un paràmetre 100% eficaç.
- Webs ubicades a servidors de Catalunya: hi ha diversos mecanismes que permeten obtenir la ubicació d'una direcció IP, i a la inversa, com per exemple els organismes RIPE[59] o ESNIC[60], i també empreses com IP2Location[61]. A partir d'aquestes dades es poden emprendre accions que permetin al *crawler* seguir les direccions IP obtingudes.
- Webs en llengua catalana: el fet que una pàgina estigui escrita en català ja és indicatiu de pertinença a la bibliografia catalana. La llengua és un factor clau per poder realitzar la compilació de recursos, ja que és un tret diferencial respecte la resta de recursos. Aquesta estratègia resulta molt eficaç en projectes com el suec per la singularitat de la seva llengua respecte les altres. En el cas de la llengua catalana també es poden trobar aspectes que puguin ser identificats per un robot prèviament programat[62].

---

[58] En el moment que es va presentar *la Memòria del plantejament del projecte PADICAT* (desembre de 2005), el domini .cat encara no havia rebut l'aprovació per part de l'ICANN, per tant no es coneixien estimacions de l'ús que se'n faria. Actualment, amb dades de principi de 2007, s'han registrat més de 21.000 dominis amb l'extensió .cat.

[59] RIPE (Réseaux IP Européens) és la secció europea de IANA (Internet Assigned Nunmbers Authority). **Vegeu**: RIPE. *Network coordination center* [En línia]. Amsterdam: RIPE, 2006. [Data de consulta: 12/08/2006]. Disponible a: <http://www.ripe.net/>.

[60] ESNIC és l'agència espanyola que gestiona els dominis dins el marc del Network Information Center. **Vegeu**: ESNIC. *Registro de dominios ".es"* [En línia]. Madrid: Ministerio de Industria, Turismo y Comercio, 2006. [Data de consulta: 12/08/2006]. Disponible a: <https://www.nic.es/>.

[61] IP2Location és una empresa que es dedica a oferir dades sobre les localitzacions de les Ips, dominis o proveïdors de serveis d'Internet. **Vegeu**: IP2Location. *IP Address to Country, Region, City, Latitude, Longitude, ZIP Code, Internet Service Provider (ISP) and Domain Name* [En línia]. Bradenton: IP2Location, octubre de 2006. [Data de consulta: 12/08/2006]. Disponible a: <http://www.ip2location.biz/>.

[62] El projecte PADICAT ha comptat amb la col·laboració del l'Institut d'Estudis Catalans i la Universitat Politècnica de Catalunya per fer estudis sobre freqüències de paraules i fórmules gramaticals pròpies de la llengua catalana.

Captura manual

El sistema admet la possibilitat d'introduir manualment URLs al robot amb la finalitat de ser capturades, en el cas que aquestes no hagin estat detectades en la recopilació automàtica.

Captura a través del dipòsit voluntari

PADICAT contempla la possibilitat de la creació d'una plataforma que permeti el dipòsit voluntari dels recursos als mateixos productors. Aquesta estratègia de captura funcionarà previ acord amb l'autor del recurs, el qual mitjançant un sistema de transmissió de dades (ja sigui via ftp o a través d'un *crawler*) podrà enviar els documents de forma periòdica o segons les actualitzacions que es realitzin.

Captura focalitzada de recursos digitals relacionats amb esdeveniments

Aquesta estratègia de captura s'explicarà de forma més detallada en els següents apartats ja que representa el conjunt de treball del projecte TematiCAT.

3.3.2.2. Tractament dels recursos

Una vegada capturats els recursos es fa un complex procés d'organització que té com a finalitat la correcte gestió de la col·lecció, i la garantia de la posterior recuperació i preservació. Aquests procés el porten a terme un equip d'editors que s'encarreguen de les tasques d'identificació, descripció amb metadades, emmagatzematge i preservació dels recursos.

Identificació permanent dels recursos

Donat que els recursos són objectes digitals i aquests estaran allotjats en un entorn en xarxa, es farà servir el sistema URI (Uniform Resource Identifier), el qual està composat per tres elements que tenen la funció de mantenir la informació sobre la localització, denominació i descripció:

- URL (Uniform Resource Location), és la cadena de caràcters amb la qual s'assigna una direcció única a cadascun dels recursos d'informació disponibles a Internet.
- URN (Uniform Resource Name), caràcters destinats a donar un nom únic i permanent a un determinat recurs.
- URC (Uniform Resource Característic), és el sistema que permet incrustar metadades al recurs.

Metadades

L'assignació de metadades és una acció clau per poder assegurar la recuperació dels recursos mitjançant la seva descripció. A part de la recuperació, s'ha de tenir en compte que la descripció dels documents permet la seva inclusió en catàlegs, i per tant fer-los accessibles a través de més punts d'accés.

El projecte PADICAT a partir d'un estudi dels models de descripció mitjançant metadades aposta per l'estàndard Dublin Core, el qual ja va ser adoptat amb anterioritat pel Consorci de Biblioteques Universitàries de Catalunya (CBUC). Dublin Core presenta diferents nivells de metadades, segons l'exhaustivitat de descripció, que es pot aplicar en funció de la tipologia recurs.

Algunes de les metadades, com la URL, títol o el tipus de format de l'arxiu es poden extreure de forma automàtica mitjançant programaris d'indexació. La resta de metadades, les quals no queden definides en la *Memòria del plantejament* (desembre 2005), són donades pels agents editors del projecte PADICAT. Tot i així, es pot avançar quina seria la configuració d'un nivell mínim de descripció mitjançant Dublin Core:

Metadades	Especificacions
Contribuïdor / Autor	*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"
Títol	*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*
Data de creació	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*
Data de captura	**Data de 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 data de publicació i/o distribució del document. Doneu obligatòriament un any aproximat. Si el mes i el dia no el coneixeu no cal que en doneu cap d'aproximat ▪ *Si no es dóna una data de publicació i/o distribució, el sistema dóna per defecte la data d'introducció del document en el dipòsit*

Llengua	**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
Matèries	**Termes extrets d'un vocabulari local controlat que defineixen el contingut del document** ▪ Doneu aquests termes a partir de l'índex de matèries 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 ▪ En el cas d'utilitzar subencapçalaments separeu-los amb un guió llarg (--) ▪ Doneu la primera lletra de cada concepte en majúscula Si el document tracta diferents temes repetiu l'element (cada tema en una casella diferent)
Resum	**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)

Fig. 13. Camps de descripció Dublin Core

Emmagatzematge

Els recursos capturats, una vegada han passat pel procés de tractament, són emmagatzemats en format comprimit (ARC) a un dipòsit que n'assegura l'accés en tot moment. En una primera fase del projecte (2006-2008) aquest dipòsit tindrà una capacitat de 10 Tb, i està previst que el sistema sigui de doble còpia, és a dir, que la informació s'allotjarà en dos dipòsits diferents. El motiu que la col·lecció estigui duplicada no és per raons d'accessibilitat de la informació a través d'una arquitectura distribuïda, sinó que es contempla com una còpia de seguretat.

Preservació

Tenint en compte que el projecte PADICAT està en la seva fase inicial, encara no s'ha fixat cap política de preservació concreta, tot i que es contempla que les estratègies més vàlides en aquest àmbit poden ser la migració periòdica o l'emulació del programari i maquinari, ambdues tenint com a finalitat assegurar l'accés als recursos i mantenir la seva aparença (*look and feel*). Un dels principals obstacles que es preveu en l'elaboració d'una política de preservació és la necessitat de manipulació dels recursos, la qual no està emparada per un text legal.

La primera aproximació que es va realitzar en el terreny de la preservació va ser un estudi d'una mostra d'arxius provinents d'Internet, per fer una estimació de les tipologies que haurà de gestionar el sistema. Aquest estudi va mostrar que gran part dels arxius que són objecte de captura corresponen a formats estàndards, que poden simplificar la tasca de preservació. Fent ús de dades absolutes, l'estudi reflecteix que d'una mostra de 700.000 arxius un 97% són formats estàndards, que desglossats

corresponen a un 53% de formats html/txt, un 21% de formats d'imatge jpg o gif i un 3% són format pdf.

3.3.2.3. Accés als recursos

Com s'ha esmentat en l'anterior apartat, no existeix a Espanya un text legal que tracti la matèria dels recursos electrònics. Aquesta mancança afecta els processos de captura, però sobretot el de manipulació (sempre que sigui necessari per la seva preservació) i el de difusió. En el cas d'aquest últim queda palès en la *Memòria del plantejament* que l'accés als recursos, en línia i en obert, està limitat a allò que es recomani en els serveis jurídics de la Biblioteca de Catalunya a tal efecte[63]. Part d'aquest problemes d'accés es solucionaran a través dels acords als que s'arribi amb els productors dels recursos, però en tot cas, aquesta mesura no es podrà prendre amb tots els recursos. En última instància, en el cas de no poder oferir els recursos en línia, es pot donar accés als recursos des de les mateixes dependències de la Biblioteca de Catalunya, tal com s'ha realitzat en altres projectes a nivell europeu com el Kulturarw3 o el Netarkivet.

Tot i les reserves envers la legalitat d'oferir en obert la col·lecció de recursos del PADICAT, l'octubre de 2006 es posava en funcionament la pàgina web del projecte amb un accés total al seu dipòsit. Segons Ciro Llueca, coordinador del PADICAT, la política d'accessibilitat que es portarà terme, seguint l'exemple de molts altres projectes que s'han trobat amb el mateix obstacle legal, serà la d'oferir tots els continguts en obert, sostenint la possibilitat d'ocultar algun recurs en cas que l'autor ho demani explícitament.

Pel que fa a la part tècnica, en el procés d'accés als recursos intervenen dos programes que fan possible la visualització. Per una part tenim Nutchwax Search Machine que s'encarrega de cercar els recursos dins la col·lecció en funció dels criteris establerts per l'usuari, i enviar la resposta que és visualitzada en la interfície. Aquesta interfície es genera a través del programa Wera, que 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 de recurs.

[63] BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT*. Op. cit.

En els cas del web del PADICAT s'ha dissenyat una interfície amb un nivell de cerca simple i un altre de cerca avançada. En el primer únicament es poden realitzar consultes a través del text lliure, mentre que en el segon es poden fer també consultes a text lliure, amb la possibilitat d'establir limitacions en els camps del domini, format del recurs, data de publicació, a més de permetre la configuració de l'ordre de visualització dels resultats.

Fig. 14. Interfície de cerca del PADICAT

4. Disseny

A partir de les dades recavades en l'anàlisi dels projectes, i en funció de l'abast i funcionalitats del projecte TematiCAT, es procedirà en aquest apartat a plasmar quin serà el disseny del sistema d'informació amb el qual es recopilaran, es tractaran i es faran accessibles els recursos que formaran part de les col·leccions temàtiques. Per tal de minimitzar la complexitat a l'hora d'expressar el disseny de l'arquitectura del sistema, l'apartat s'ha dividit en tres punts coincidents amb els *subsistemes* de selecció i captura, de tractament i de recuperació i visualització. En cadascun d'aquests punts es descriuran els processos que hauran de fer els mòduls, i s'establiran quins són els requisits de les aplicacions que els hauran de portar a terme. Finalment es realitzarà un *benchmarking* d'aplicacions per tal de tenir les dades suficients per poder triar les més convenients per les característiques del projecte.

4.1. Disseny del sistema d'ingestió de recursos web

El sistema d'ingestió de recursos s'encarregarà de realitzar els processos de selecció i captura dels objectes digitals que formaran part dels dossiers temàtics. A diferència dels processos de captura integral o selectiva, la recopilació de webs relacionats amb esdeveniments és més delicada donat al caràcter efímer de la informació. Tenint en compte la duració del procés de captura de recursos d'Internet[64] i en funció de la transcendència de l'esdeveniment, s'hauran de traçar diferents estratègies d'ingestió de webs per tal de perdre el mínim d'informació d'interès possible.

Basant-nos en l'experiència d'altres projectes, podem definir dos tipus d'esdeveniments d'interès, els programats i els fortuïts (no programats). Un esdeveniment programat podria ser, per exemple, la celebració d'unes eleccions, les quals es donen a conèixer amb un ampli període de temps i generen informació a priori i a posteriori de la seva realització. En canvi, els esdeveniments fortuïts es defineixen per la immediatesa i inconcreció, i podria tractar-se, per exemple, d'una catàstrofe o un fet que salta a la primera línia de notícies d'interès de forma inesperada.

[64] A partir de les dades obtingudes en les diferents captures realitzades en el projecte PADICAT durant l'any 2006, es sap que la captura de 25.000 URLs (100 Gb) del domini web català pot arribar a durar fins a 15 dies, i que la indexació d'aquests recursos pot requerir fins a 30 dies.

Per explicar més detalladament la diferència entre aquest dos tipus d'esdeveniments, podem prendre com exemple dos casos del projecte Minerva. Per una banda, representant els esdeveniments programats, podem observar la recopilació de webs en relació a les eleccions al Congrés de l'any 2002[65]. Per l'altra banda, com exemple d'un esdeveniment fortuït, podem trobar el recull de webs que es va fer en motiu dels tràgics fets de l'atemptat terrorista a Estats Units del setembre de 2001[66].

L'estratègia de selecció i captura de recursos vindrà donada en funció del tipus l'esdeveniment que es vulgui cobrir. En el cas dels esdeveniments programats existirà la possibilitat de plantejar una preselecció de recursos més acurada, concreta, i amb una inversió de recursos tecnològics sensiblement inferior, mentre que amb els esdeveniments fortuïts s'haurà d'executar una captura directament a Internet, mitjançant l'ús d'algoritmes de cerca simples[67]. Els processos de recopilació dels recursos s'explica amb més detall en el següent apartat dedicat a les estratègies de selecció i captura segons la tipologia d'esdeveniments esmentats.

El recursos web capturats seran emmagatzemats en format comprimit (ARC) en un dipòsit propi per aquest projecte. Aquest dipòsit estarà format per dues particions, una dedicada a l'allotjament definitiu dels recursos i una altra que s'utilitzarà com un espai diferenciat de treball, on es realitzaran els processos de tractament i validació dels arxius que seran preservats.

Resumint els processos mencionats, les funcionalitats que haurà de realitzar el sistema d'ingestió de recursos són les següents:

- Captura de recursos web d'Internet a través d'URLs i algoritmes simples
- Captura de recursos web del dipòsit del PADICAT mitjançant un programa que en permeti la recuperació

---

[65] MINERVA PROJECT. *107th Congress Web Archive* [En línia]. Washington: Library of Congress, octubre de 2006. [Data de consulta: 15/01/2007]. Disponible a: <http://lcweb2.loc.gov/cocoon/minerva/html /107th/ search.html>.

[66] MINERVA PROJECT. *September 11 Web Archive* [En línia]. Washington: Library of Congress, octubre de 2006. [Data de consulta: 15/01/2007]. Disponible a: <http://lcweb2.loc.gov/cocoon/minerva/html/sept11 /sept11-about.html>.

[67] En un escenari ideal, la selecció i captura temàtica es podria fer integrament a partir d'Internet, mitjançant el que s'anomena *focused web crawling*, fent servir programes basats en complexes algoritmes que seleccionen i recopilen els webs en funció d'un tòpic. Tot i que aquestes tècniques de cerca avançada ja han estat posades a prova amb col·leccions digitals, cap dels projectes analitzats que segueixen el model selectiu o híbrid fa ús encara d'aquest sistema avançat de captura de recursos. **Vegeu***:* ESTER, MARTIN; GROSS, MATTHIAS; KRIEGEL, HANS-PETER. "Focused Web Crawling: [En línia] a generic framework for specifying the user interest and for adaptive crawling strategies". *Twenty-Seventh International Conference on Very Large Databases*, 2001. [Data de consulta: 25/01/2007]. Disponible a: <http://citeseer.ist.psu.edu/ester01focused.html>.

4.1.1. Estratègies de captura selectiva de recursos digitals

Abans entrar a definir les estratègies de cerca en cadascun del tipus d'esdeveniments anteriorment citats, s'explicarà breument i de forma genèrica en que consisteix cada tipus de cerca, ja que aquestes es poden combinar en les diferents estratègies segons uns interessos predeterminats.

- Recuperació a través d'URL: aquest tipus de recuperació es pot aplicar tan a Internet com directament al dipòsit del PADICAT. Consisteix en la introducció manual d'URLs amb la finalitat de recuperar un recurs en concret. En el cas de la recuperació a Internet es farà servir un *crawler*, al qual se li podran configurar els paràmetres de captura del web. En el cas dels recursos del dipòsit del PADICAT es farà servir un mòdul de consulta a través del camp de descripció URL.

- Cerca per paraula clau basada en el fitxer invers: aquest tipus de cerca és aplicable als arxius del dipòsit PADICAT, els quals han estat indexats i han generat un fitxer invers. Mitjançant l'agent de consulta es pot executar una cerca a text lliure, amb la possibilitat de crear equacions més complexes a partir dels operadors booleans.

- Cerca basada en descriptors de matèria: també és únicament aplicable als arxius del dipòsit del PADICAT. En el projecte PADICAT està previst que agents humans facin una breu descripció dels recursos a través de descriptors. Novament, mitjançant l'agent de consulta, es podrà interrogar el camp referent als descriptors per tal de recuperar recursos sobre determinades matèries.

- Cerca basada en paràmetres de filtres condicionals: encara que pot ser aplicable als dos casos, aquest tipus de cerca ens interessa sobretot a nivell de la recuperació de recursos d'Internet. La majoria de *crawlers* a gran escala permeten la configuració de paràmetres per filtrar la captura de webs. En aquest cas la captura vindria determinada pel compliment de certs paràmetres configurats per valors. És a dir, que el recurs només seria capturat en el cas que hi hagués un valor en concret (que seria un o més termes) en un determinat lloc del document (per la riquesa d'informació serien interessants els títols dels documents o els encapçalaments). En funció del *crawler*, aquest tipus de cerca podria requerir un desenvolupament del programa per adaptar-lo a les necessitats.

- Cerca basada en un algoritme vectorial: aquest tipus de cerca requereix per una banda un considerable desenvolupament informàtic del programa per tal de poder executar aquest tipus de consulta, i per altra banda, creació de plantejament lògics per tal d'elaborar un model vectorial. A grans trets, aquest model es basa en la representació de la freqüència d'aparició dels termes i la distància que hi ha entre aquests. Inicialment es parteix d'un document ideal, del qual es calculen el seus vectors per tal de buscar altres documents similars. Aquest mètode no només recupera documents, sinó que en calcula la rellevància en funció de la similitud amb el model ideal donat.

Tal com s'ha esmentat en l'anterior apartat, hi ha programes que estan preparats per poder implementat mòduls per realitzar captures focalitzades del web en funció de tòpics o estructures lingüístiques. Es considera que posar en pràctica aquests mètodes de captura podria resultar molt arriscat en aquest projecte, ja sigui perquè requereix una infraestructura tecnològica important, o bé per la manca d'experiències conegudes. Per aquest motiu es planteja la possibilitat de fer ús de les tècniques focalitzades de captura en una següent fase del projecte.

4.1.1.1. Estratègia de captura per esdeveniments programats

En el cas dels esdeveniments programats es compte amb un temps inicial que s'ha de destinar a fer un plantejament de la selecció i recopilació que ajudi a minimitzar els alts costos en temps i recursos que significa una captura a Internet. Per aquest motiu es proposa una estratègia basada en una preselecció de recursos relacionats amb el fet al dipòsit del PADICAT, ja que aquests han passat per un filtre que garanteix que formen part de la producció digital catalana, i a més han estat indexats i catalogats.

La preselecció es farà amb l'agent de consulta, el qual permetrà realitzar interrogacions als camps de descripció: URL, matèria i al fitxer invers. Està previst que quan es realitzi la cerca, es limiti a la recuperació de webs entrats en el dipòsit els dos últims anys, ja que en molts casos, quan llavors volguéssim capturar el web actual, ens podríem trobar amb recursos que han canviat la seva adreça o senzillament que han desaparegut.

En funció del tema de l'esdeveniment i la seva singularitat els resultats de la preselecció poden ser més o menys satisfactoris. Així doncs, si per exemple s'està fent una preselecció per crear una col·lecció temàtica sobre la celebració d'unes eleccions a Catalunya o la consecució d'un títol esportiu per part d'un club català, el resultat serà

bastant positiu, ja que en el dipòsit es podran localitzar fàcilment webs relacionats amb temes de política o esports respectivament. En canvi, en el cas de voler crear una col·lecció temàtica sobre un esdeveniment molt concret i particular, pot resultar que el dipòsit no posseeixi recursos que cobreixin aquest tema. En casos molt evidents de manca d'utilitat de la preselecció, es pot optar per passar a l'estratègia de recuperació directa a Internet.

Una vegada preseleccionats els recursos, aquest seran allotjats al dipòsit del TematiCAT i passaran per la fase de validació, aspecte que es comentarà més detalladament en el seu corresponent apartat. Posteriorment es crearà una llista d'URLs dels elements preseleccionats i s'establirà una freqüència de captura automàtica de cada recurs a Internet.

En funció del grau de satisfacció de la preselecció, que vindrà donada pel nombre de documents identificats al dipòsit, es continuarà, o no, la cerca de nous recursos a Internet. Aquesta segona estratègia de cerca es realitzarà utilitzant diferents mètodes. Per una banda es configuraran els paràmetres corresponents perquè el *crawler* segueixi els enllaços incrustats en les pàgines web, ja que aquests solen tenir relació temàtica. Per altra banda es llançaran altres robots configurats amb filtres condicionals per localitzar nous recursos. Per evitar la duplicitat de tasques, s'usarà la mateixa llista d'URLs de webs preseleccionades, i es configuraran els robots perquè no visitin aquest recursos, ja que s'estaran executant a l'hora els treballs que s'encarreguen de fer el seguiment periòdic.

Els nous webs seleccionats hauran de ser validats, i com amb els preseleccionats, s'establirà una freqüència de captura.

4.1.1.2. Estratègia de captura per esdeveniments no programats

En el cas dels esdeveniments no programats, si es creu que hi ha alguna possibilitat d'èxit, es pot intentar fer una preselecció de recursos al dipòsit del PADICAT i continuar amb el procediment de l'anterior estratègia. En tot cas, el protocol a seguir serà un altre, que tindrà com a font prioritària de recursos Internet.

La captura dels recursos es realitzarà a través d'un *crawler* que serà configurat amb diversos paràmetres, en funció de les necessitats de cada cerca. Perquè el *crawler* comenci a navegar se li han de donar unes determinades coordenades, que són les URLs. Per tal de centrar-nos directament en la producció digital catalana, pressuposarem que aquesta es troba integrament recollida dins el dipòsit del

PADICAT[68]. Així doncs partirem d'un llistat de dominis extret de la base de dades del PADICAT per tal de començar a executar la cerca.

En aquesta estratègia de cerca es pot utilitzar el mètode de cerca basat en filtres condicionals o el basat en el model vectorial, el qual s'escollirà en funció de les necessitats de cada cerca temàtica. Tot i així, es prioritzarà la cerca basada en filtres condicionals, ja que la programació d'un model vectorial pot resultar molt costosa. En tot cas, abans d'executar una cerca definitiva, es farà una prova pilot amb una mostra per tal de pronosticar el seu grau d'èxit, i modificar l'estratègia en cas que sigui necessari.

Els webs que compleixin els criteris seran capturats, i posteriorment passaran la fase de validació, on seran indexats i descrits per tal de ser allotjats en el dipòsit i establir la seva freqüència de captura.

4.1.2. Característiques dels recursos digitals del dipòsit

Per tal de facilitar el procés de captura, tractament i el posterior accés als recursos digitals s'analitzaran els resultats obtinguts en algunes de les captures realitzades pel projecte PADICAT, amb la finalitat recavar dades sobre la tipologia de formats i el volum dels arxius. Aquest anàlisi ens ha de permetre obtenir la informació necessària per poder fer una estimació dels requeriments dels programes que realitzaran els processos de captura, i poder indagar sobre la parametrització del procés. Al mateix temps, aquest estudi servirà per poder tenir més dades a l'hora de definir la política de preservació dels documents.

La mostra d'arxius que és subjecte d'anàlisi es va capturar l'any 2005[69] amb la finalitat de realitzar una prova pilot del sistema d'ingestió de recursos. El volum de la mostra va ser de 25.000 seus web, composats per més de 2,5 milions d'arxius i amb un volum total de 100 Gb. Els resultats mostren que el gruix dels arxius són formats estàndards, dels quals un 56% són arxius text/html i un 37% són formats d'imatge (19% jpeg, 15% gif i un 3% pdf).

[68] L'experiència dels projectes estudiats ens han demostrat la dificultat de recopilar els recursos web d'un determinat territori geogràfic i polític. Tot i que el cas català té el factor diferencial de la llengua, els estudis que es van realitzar en el projecte PADICAT amb estructures lingüístiques no van ser del tot exitosos.

[69] S'ha de tenir en compte que aquests estudi és del 2005, i que per tant les dades sobre el nombre d'arxius per format i el seu tamany avui en dia poden haver variat sensiblement. Els últims anys s'ha estès la participació per part del internautes mitjançant aplicacions com els blocs, entre d'altres, que han possibilitat que la xarxa fos un espai idoni per poder compartir imatges, tant fixes com en moviment.

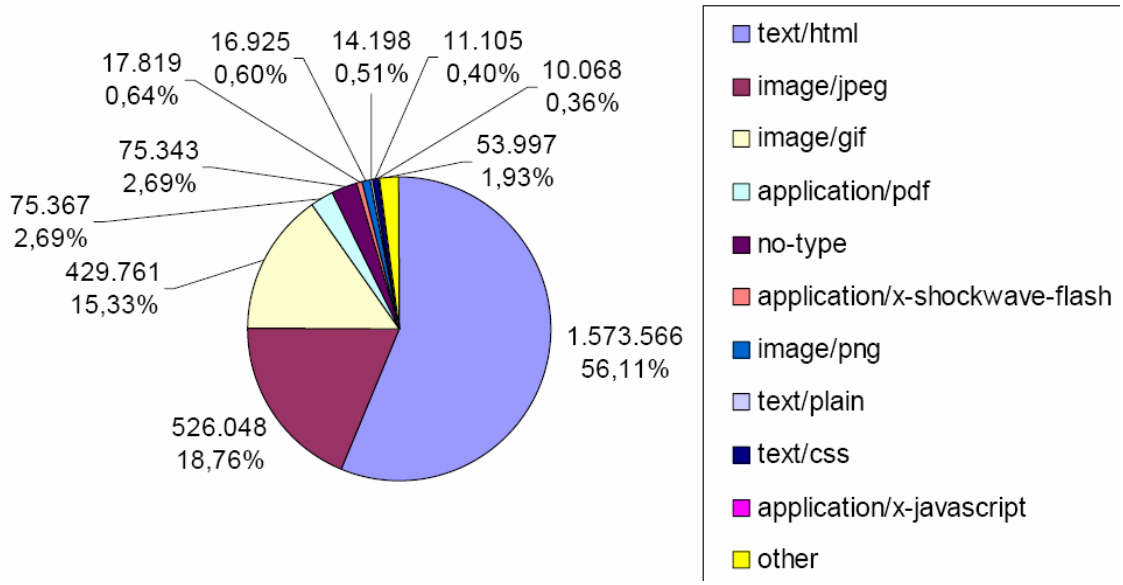


Fig. 15. Estudi de formats

Pel que fa el volum dels arxius, la mostra capturada reflecteix que gran part d'aquests, un 68%, tenen un pes relativament petit (menys d'1 Mb), tot seguit dels mitjans (d'1 a 100 Mb) amb un 30% del total, i finalment els grans (més de 100 Mb) que únicament tenen una representació aproximada del 2%.

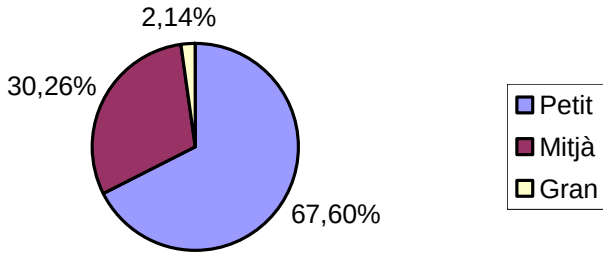


Fig. 16. Volum dels arxius

En relació al volum de la mostra, un 25% del volum està representat per documents pdf, tot seguit d'un 21% d'arxius en format text/html i un 15% en format jpeg.

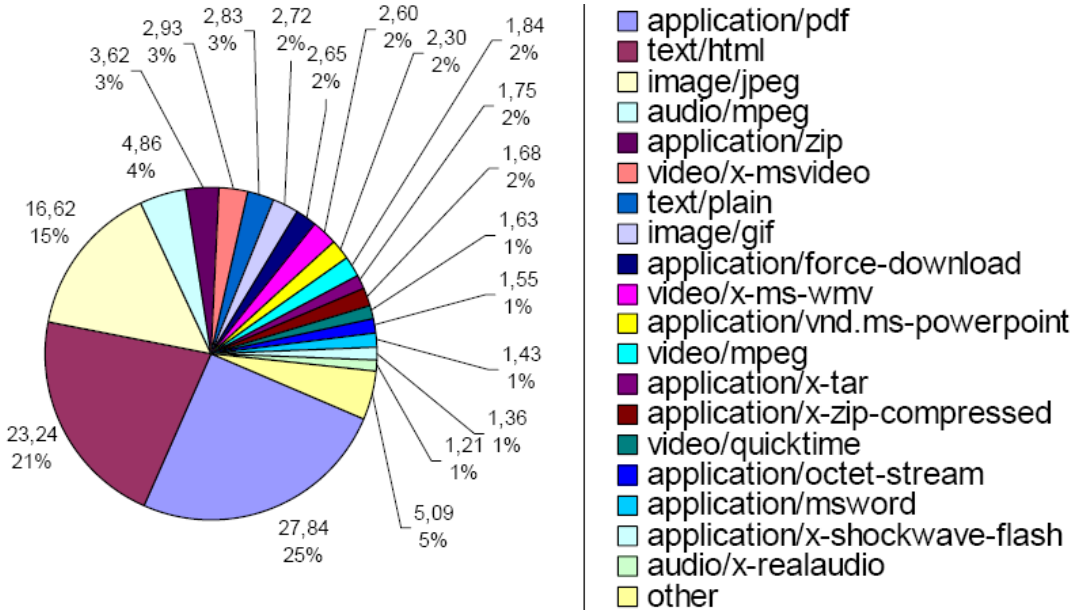


Fig. 17. Relació de volum de formats

En els resultats de la captura del PADICAT no donen xifres sobre el nombre de recursos estàtics i dinàmics que formen part de la mostra. Tot i així, fent referència a altres projectes estudiats en apartats anteriors, si que es coneix que els webs dinàmics no suposen cap complexitat que no es pugui solucionar amb la parametrització del *crawler*, llevat de tres casos: els webs que tinguin bases de dades en obert, webs amb algun tipus de clau d'accés, i els webs amb calendari accionat a través de JavaScript, no seran capturats per tal d'evitar errors fatals. En el cas dels webs amb calendaris, s'impedeix la seva descarrega perquè normalment aquests no tenen una data límit, i quan el *crawler* els intenta capturar, entren en un bucle que pot ocasionar problemes al servidor.

4.1.3. Requeriments tecnològics del programari de captura de recursos a Internet

El programa de captura haurà de complir els següents requeriments per tal de poder ser adaptat a la plataforma tecnològica del projecte TematiCAT:

A nivell d'implementació

- Programari de codi obert per tal de ser desenvolupat posteriorment en cas que sigui necessari. Els òrgans de l'administració pública recomanen fer extensiu l'ús de programari lliure
- Compatibilitat amb el sistema operatiu Linux: com en el cas del programa que es vol adquirir, el sistema operatiu que suportarà tota la plataforma tecnològica serà de codi obert.
- Compatibilitat amb el servidor de pàgines web Apache 2.0
- Compatibilitat amb el servidor d'aplicacions web Tomcat 5.0

A nivell de funcionalitats

- Capacitat de captura integral: possibilitat de programar una captura a partir d'un cert nombre d'URLs, en la qual el *crawler* continuï el procés amb expansió viral a través dels enllaços dels webs
- Capacitat de captura selectiva: possibilitat de programar una captura de determinats webs
- Capacitat de captura per domini: possibilitat de programar una captura sota el criteri del domini. Aquest requeriment és molt important per tal de poder aprofitar l'existència del domini .cat
- Capacitat de captura per IP: possibilitat de programar una captura dirigida a una determinada direcció IP, que pot fer referència a un editor amb el qual s'ha arribat a un acord
- Capacitat d'emmagatzematge amb compressió: els recursos capturats hauran de ser allotjats en el dipòsit en format comprimit per tal d'economitzar l'espai virtual
- Capacitat d'automatització de períodes de captura: possibilitat de programar diferents freqüències automàticament en funció de la necessitat de cada recurs

- Capacitat de discriminar la captura de webs duplicades, perquè ja existeix una versió en el dipòsit
- Capacitat de regulació de la descàrrega d'arxius: per tal d'evitar la sobresaturació dels servidors dels clients, és important que la descàrrega dels webs es pugui controlar

4.1.4. Estudi de programes de captura de recursos web

Heritrix

Heritrix és fins avui el robot en codi obert de captura de recursos digitals més estès entre els projectes actius. Aquest programa ha estat desenvolupat en les seves últimes versions per 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.

Aquest és un programa al que se li donen unes coordenades de navegació dins Internet, i que d'acord amb una sèrie de criteris es dedica a descarregar les pàgines web que ha visitat. Hi ha diverses formes de donar les coordenades a Heritrix[70], que poden anar des de la introducció d'una única URL a seguir, fins a la descàrrega d'un domini sencer.

Heritrix està format per diversos mòduls que gestionen els diferents processos que s'executen. Inicialment hi ha el mòdul d'entrada o ingestió, al qual se li dóna una o més URLs com a coordenada. Durant la captura existeixen diversos mòduls interns que s'encarreguen de processar aquesta acció. Per exemple, la llista d'entrada passa sempre per un mòdul anomenat frontera (*frontier*), on s'analitza que tots els ítems compleixen els criteris de captura donats.

Una vegada capturat el recurs, aquest passa per un cicle de processadors, tal com podem veure en el gràfic inferior, que s'encarreguen de l'extracció d'enllaços per tal de descobrir nous recursos que es troben incrustats a les pàgines recopilades. Aquest nous recursos són agregats a la llista d'URLs per tornar a començar el cicle. El pas següent és la compressió dels arxius en format ARC[71] per tal de ser emmagatzemats.

---

[70] INTERNET ARCHIVE, ET AL. *Heritrix User Manual* [En línia]. Internet Archive, 2005. [Data de consulta: 23/01/2007]. Disponible a. <http://crawler.archive.org/articles/user_manual/index.html>.

[71] El format ARC permet que els arxius siguin indexats, els fa cercables i en proporciona una bona visualització.

Heritrix es caracteritza per tenir una arquitectura molt flexible que permet parametritzar detalladament el perfil de les captures mitjançant la implementació dels mòduls. Aquests ofereixen la possibilitat de crear treballs de captura (*jobs*) completament personalitzats en funció de les necessitats. Així doncs, es permet la limitació de la profunditat de captura, el temps de descàrrega, el nombre de documents i bits, així com el temps entre una captura i una altra per tal d'evitar la saturació del servidor client. Heritrix també permet configurar l'opció de robots.txt per tal de respectar els editors de les webs que no permeten descarregar el seu web.

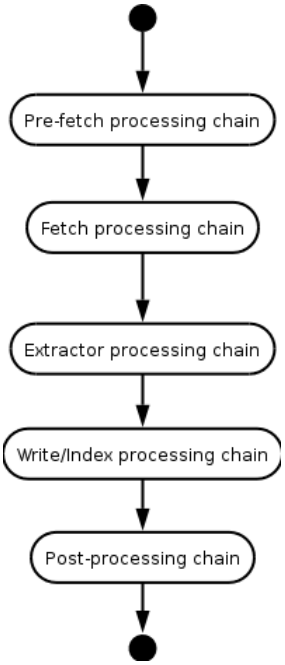


Fig. 18. Cadena de processos del programa Heritrix

En relació a la captura periòdica de webs, Heritrix permet implementar un nou mòdul, anomenat DeDuplicator[72], que permet rebutjar la captura d'un recurs, si aquest ja estat capturat anteriorment. Aquest mòdul funciona a través d'un processador, que mitjançant la indexació realitzada dels recursos, és capaç de detectar si una pàgina ha estat actualitzada.

[72] SIGUROSSON, KRISTINN. *Managing duplicates across sequential crawls* [En línia]. Reykjavík: National and University Library of Iceland, 2007. [Data de consulta: 05/06/2007]. Disponible a: <http://vefsofnun.bok.hi.is/upload/3/ManagingDuplicatesAcrossSequentialCrawls.pdf>.

HTTrack website copier

HTTrack és un programa desenvolupat en codi obert que permet la descàrrega de recursos del World Wide Web a un directori local, emulant l'estructura original del dels arxius en línia. Aquest programa és usat per Netarkivet en la fase de captura selectiva, i les referències sobre el seu funcionament són correctes.

HTTrack[73] es caracteritza per la seva facilitat d'ús i adaptació en qualsevol entorn tecnològic. Respecte el programa anterior disminueix considerablement la complexitat del procés de captura, ja que l'arquitectura d'aquest programa és molt més simple, i per tant es perden algunes de les funcionalitats que ofereix Heritrix.

Malgrat algunes diferències en les opcions avançades, HTTrack segueix el mateix mecanisme de captura que Heritrix. Es parteix d'un llistat d'URLs amb l'objectiu de ser descarregades, però en el cas d'aquesta aplicació, el robot no continua navegant pels enllaços incrustats en les pàgines web. En les opcions de descàrrega dels objectes digitals destaca la possibilitat de limitar segons el tipus d'extensió (per exemple, baixar, o no, només arxius zip) i la possibilitat d'ordenar-li la creació d'un fitxer invers de les paraules significants dels arxius descarregats.

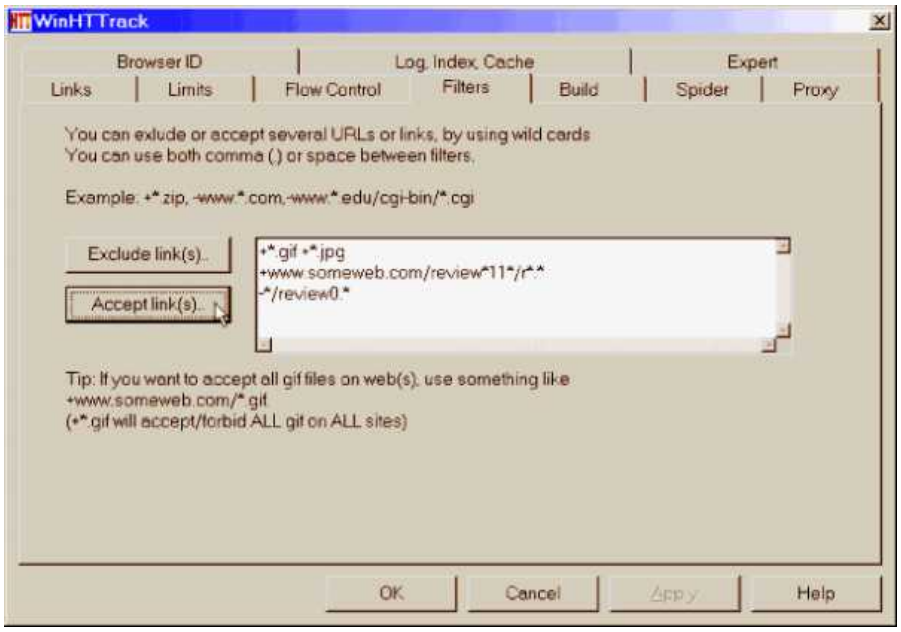


Fig. 19. Paràmetres de captura de l'aplicació HTTrack

[73] ROCHE, XAVIER, ET AL. *HTTrack Web Copier* [En línia]. França: Leto Kauler, 2007. [Data de consulta: 26/01/2007]. Disponible a: <http://www.httrack.com/page/1/en/index.html>.

Com en el programa Heritrix, HTTrack també ofereix les opcions de limitar el volum de descàrrega mitjançant la configuració dels paràmetres de profunditat de captura, nombre i volum dels recursos, i també el temps i número de connexions a un lloc web.

Fig. 20. Monitorització de procés de captura de l'aplicació HTTrack

Finalment cal destacar dues característiques de HTTrack, una positiva i una altra negativa en relació als requeriments a complir. Per la banda negativa, aquest programa no ofereix la possibilitat automàtica de compressió dels arxius quan aquests són descarregats, ja que fa un replica exacte (*mirror*) dels recurs en línia a un directori local. Com a característica positiva destaca la possibilitat d'actualitzar de forma automàtica els webs que han estat baixats.

Nedlib Harvester

Amb la posada en funcionament del projecte suec Kulturarw3 l'any 1996, es comença a desenvolupar el programa Nedlib (Networked European Deposit Library)[74], i que més endavant seria desenvolupat per el Nordic Web Archive sota el lideratge de la Helsinki University Library. Nombrosos projectes dedicats a la recopilació de llocs web han fet servir en algun moment, com han estat els casos del Netarkivet (Dinamarca), AOLA (Àustria), WebArchiv (República Txeca), entre d'altres. Com els altres *crawlers* analitzats, Nedlib està programat en codi obert, i treballa conjuntament amb una base de dades MySQL relacional.

[74] HAKALA, JUHA. "The Nedlib harvester" [En línia].*NEDLIB workshop 2000*. La Haya: desembre del 2000. [Data de consulta: 26/01/2007]. Disponible a: <http://nedlib.kb.nl/workshop/NEDLIB%20harvester.ppt#1>.

Comparat amb altres *harvesters*, Nedlib té algunes funcionalitats especials. La més òbvia és un mòdul arxiu, que té la funció de generar metadades dels recursos capturats i processar-los de manera que aquests puguin ser allotjats de forma que faciliti la indexació, mitjançant el protocol MD5 checksum, el qual actua com identificador únic de cada arxiu.

Pel que fa a la funció de captura hem pogut constatar que aquest mòdul és encara bastant inestable. A partir de la lectura d'informes de diferents projectes[75] detectem que el programa requereix un desenvolupament particular per cada plataforma on es vol implementar.

Les característiques del procés de captura són bastant semblants als anteriors programes presentats. Es parteix d'una llista d'URLs, les quals són navegades, i partir dels enllaços de les pàgines es generen noves llistes de captura. Hi ha també la possibilitat de fer restriccions a nivell de domini, *hosts*, etc., i es permet la monitorització dels processos del robot per tal de controlar les descarregar i assegurar la integritat dels robots.

Combine

Combine és un programa en codi obert que combina les funcions de captura i indexació de recursos digitals en línia. Inicialment va ser creat per NetLab i actualment és mantingut i desenvolupat pel KnowLib Group[76] del Departament de Tecnologia de la Informació de la Lund University. Aquest programa ha estat implementat en el projecte suec Kulturarw3 i en l'austríac AOLA.

Combine[77] es mostra com un dels programes més complets pel que fa a funcionalitats i flexibilitat a l'hora de ser configurat. Com la resta de programes analitzats, posseeix un mòdul de captura accionat per un llistat d'URLs que li és donat, i la possibilitat de seguir navegant les URLs que es troba en cada pàgina web. També posseeix un mòdul d'extracció de metadades que són guardades en una base de dades SQL per tal del ser explotades.

---

75 ASCHENBRENNER, ANDREAS. "AOLA: the austrian on-line archive". *Long-term preservation of digital material -building an archive to preserve digital cultural heritage from the internet* [En línia]. Viena: Information & SoftwareEngineering Group, 2004. [Data de consulta: 26/01/2007]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/AOLA.html>.

76 KNOWLIB. *Knowledge Discovery and Digital Library Research Group* [En línia]. Lund: Department of Information Technology, Lund University, octubre de 2006. [Data de consulta: 26/01/2007]. Disponible a: <http://www.it.lth.se/knowlib/>.

77 KNOWLIB. *The Combine harvesting robot* [En lína]. Lund**:** Department of Information Technology, Lund University, gener de 2007. [Data de consulta: 26/01/2007]. Disponible a: <http://combine.it.lth.se/ #features>.

La novetat respecte la resta de *crawlers,* és que Combine treballa en el mode de captura focalitzades de webs (*web focused crawler*), basat en una classificació de tòpics i en la detecció d'elements lingüístics particulars. Aquestes aplicacions són possibles mitjançant la integració de nous mòduls que s'implementen a la base de Combine.

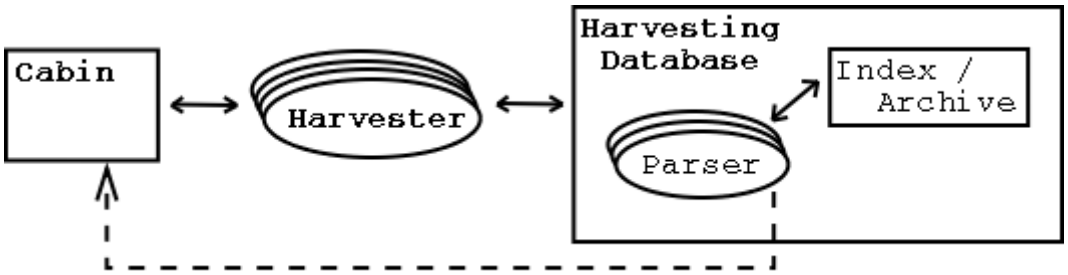


Fig. 21. Esquema dels processos de Combine

Tal com mostra el gràfic superior[78], Combine està format per tres mòduls bàsics. El mòdul anomenat *Cabin* és el controlador de tot el sistema. Aquest conté la llista d'URLs que han de ser capturades, i també les que ja ho han estat. En el mateix mòdul es configuren els paràmetres de captura per cada treball i a continuació executa el mòdul de captura, el qual pot executar diferents treballs alhora per tal d'incrementar l'índex de descàrregues. Els recursos capturats són dirigits al tercer mòdul (*harvesting database*), on són extrets els enllaços de les pàgines web, amb la finalitat de ser inclosos en noves llistes. Per altra banda, es crea un índex amb determinades dades per cada URL que és guardat en una base de dades.

De la mateixa manera que en la resta de programes, Combine permet la parametrització de la profunditat de captura del web, nombre i volum del recursos i calibratge del temps de connexió. Per la resta de funcionalitats més avançades, hi ha manual d'ús molt complet publicat en línia, que ofereix diferents exemples de programació de tòpics i filtres lingüístics[79].

[78] ASCHENBRENNER, ANDREAS. "Adapting the Combine crawler". *Long-term preservation of digital material - building an archive to preserve digital cultural heritage from the internet* [En línia]. Viena: Information & SoftwareEngineering Group, 2004. [Data de consulta: 26/01/2007]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Adapting_Combine_crawler.html>.

[79] DRAKOS, NIKOS i MOORE, ROSS. *Documentation for the Combine (focused) crawling system* [En línia]. Traducció de Anders Ardö. Lund: Department of Information Technology, Lund University, febrer de 2007. [Data de consulta: 01/03/2007]. Disponible a: <http://combine.it.lth.se/documentation/>.

4.1.5. Requeriments tecnològics del programari de selecció de recursos del dipòsit PADICAT i d'indexació

El programa encarregat de la selecció de recursos del dipòsit PADICAT i la indexació dels recursos capturats a Internet haurà de complir els següents requeriments per tal de poder ser adaptat a la plataforma tecnològica del projecte TematiCAT:

A nivell d'implementació

- Programari de codi obert per tal de ser desenvolupat posteriorment en cas que sigui necessari. Els òrgans de l'administració pública recomanen fer extensiu l'ús de programari lliure
- Compatibilitat amb el sistema operatiu Linux: com en el cas del programa que es vol adquirir, el sistema operatiu que suportarà tota la plataforma tecnològica serà de codi obert.
- Compatibilitat amb el servidor de pàgines web Apache 2.0
- Compatibilitat amb el servidor d'aplicacions web Tomcat 5.0

A nivell de funcionalitats

- Capacitat d'indexació automàtica dels recursos capturats del web amb la finalitat de crear un fitxer invers
- Capacitat d'extreure i generar metadades automàticament, com per exemple el títol del recurs, la data de captura, el seu pes, i la URL.
- Capacitat de recuperar recursos digitals allotjats en el dipòsit PADICAT a través d'una interfície de consulta de l'agent editor
- Capacitat de recuperar recursos del dipòsit TematiCAT a través d'una interfície de consulta de l'usuari final
- Capacitat de gestionar arxius en format comprimit

4.1.6. Estudi de programes de consulta i d'indexació de recursos web

Lucene

Lucene[80] és un motor de cerca de text programat en codi obert desenvolupat per la Fundació Apache. Lucene es caracteritza per ser un *software* amb un gran nombre de funcionalitats, una escalabilitat flexible i un desenvolupament estable, motiu pel qual s'ha convertit en la plataforma de nombrosos projectes.

Lucene posseeix un sistema multifunció que li permet executar tasques d'indexació i cerques simultàniament, amb uns requeriments tecnològics mínims, però que en el cas de treballar amb una plataforma informàtica professional augmenta substancialment el seu rendiment.

El mòdul d'indexació de Lucene permet programar diferents treballs alhora amb diferents paràmetres d'extracció de dades. L'objectiu final és elaborar un fitxer invers dels termes significants, els quals són introduïts en una base de dades SQL, per tal de facilitar la tasca de recuperació d'informació.

Lucene incorpora un ampli ventall de possibilitats en el mòdul de consulta mitjançant una potent sintaxis de cerca i la capacitat de configurar mètriques de rellevància per classificar els resultats. Permet la possibilitat de fer consultes a diferents camps (títol, autor, contingut, etc.) i limitar per intervals de dates. Disposa de la capacitat d'elaborar interrogacions a nivell avançat a partir de la utilització d'operadors booleans, de proximitats i de truncament, i també, la possibilitat de crear equacions de cerca mitjançant algoritmes en model vectorial. En el cas que existeixin, Lucene permet realitzar cerques en índexs múltiples, i la seva actualització *en calent*. Els resultats obtinguts es poden agrupar en llistes en funció d'un *ranking*, o bé classificar-los segons els contingut d'un camp predeterminat.

Nutch Wax

Nutch Wax[81] (Nutch Web Archive eXtensions) és un programa en codi obert que porta a terme les funcions d'indexació i cercar dins col·leccions d'arxius. Aquest programa ha estat desenvolupat per la Fundació Apache, que ha comptat amb la col·laboració d'institucions tant rellevants en el món de la preservació web com la NWA, el IIPC i

[80] APACHE SOFTWARE FUNDATION. *Apache Lucene* [En línia]. [USA]: Apache Software Fundation, 2007. [Data de consulta: 01/03/2007]. Disponible a: <http://lucene.apache.org/java/docs/>.

[81] INTERNET ARCHIVE. *Nutch Wax* [En línia]. Washington: Internet Archive, 2007. [Data de consulta: 01/03/2007]. Disponible a: <http://archive-access.sourceforge.net/projects/nutch/index.html>

Internet Archive. Nutch Wax ha estat construït sobre la base de Lucene, amb la finalitat de treballar sobre els estàndards dels mòduls d'indexació i consulta, i desenvolupar un potent motor de cerca aplicat a dipòsits de recursos web.

Aquest programa es caracteritza per una gran flexibilitat en la seva implementació ja que la seva estructura modular permet als desenvolupadors treballar amb facilitat, rendibilitzar la seva potència, i adaptant-lo a les necessitats de qualsevol projecte. L'escalabilitat de Nutch Wax permet la seva utilització en qualsevol tipus d'entorn, ja sigui en local, en una Intranet o a l'escala de tot el web.

El mòdul d'indexació s'encarrega de generar un fitxer invers a partir dels documents capturats i l'extracció automàtica de certes metadades com el títol, URL, data, tipus d'arxiu, etc.). En el marc d'aquesta funcionalitat Nutch Wax es mostra més potent que Lucene, ja que utilitza un subprograma anomenat Hadoop[82] que li permet aprofitar al màxim les possibilitats que ofereix la plataforma tecnològica. El *framework* de Hadoop proporciona un entorn de processament distribuït de dades transparent i fiable per a qualsevol aplicació. Aquest implementa el paradigma computacional conegut com a *map/reduce*, en el qual l'aplicació és dividida en diversos fragments d'execució, on cadascun dels quals pot ser executat i/o tornat a executar en qualsevol dels nodes del clúster. A més, Hadoop proporciona un sistema d'arxius distribuït que emmagatzema les dades en els diferents nodes, obtenint d'aquesta forma un ample de banda agregat molt alt. Tant el *map/reduce* com el sistema d'arxius distribuït han estat dissenyats de manera que el *framework* soluciona automàticament la possible caiguda de nodes.

Nutch Wax ofereix també una alta gamma d'eines per poder efectuar consultes en el conjunt de documents que integren la col·lecció. Com Lucene, permet la possibilitat de limitar la cerca de determinats camps, i l'ús de tècniques avançades de cerca com els operadors booleans o els algoritmes vectorial. També disposa de paràmetres de configuració dels resultats de la cerca, mitjançant l'ordenació per camps o mètodes de mètrica de rellevància dels documents.

La popularització de Nutch Wax ha portat a crear passarel·les d'estandardització amb aplicacions de visualització com Wayback Machine o Wera, evitant d'aquesta manera alts costos en desenvolupament i adaptació entre les aplicacions de consulta i accés als recursos.

[82] APACHE SOFTWARE FUNDATION. *Welcome to Hadoop!* [En línia]. [USA]: Apache Software Fundation, 2007. [Data de consulta: 01/03/2007]. Disponible a: < http://lucene.apache.org/hadoop/>.

4.2. Disseny del sistema de tractament dels recursos web capturats

Els processos de tractament dels recursos recopilats tenen lloc a l'àrea reservada de treball del dipòsit, i representen el pas previ a l'allotjament definitiu dels webs que formaran les col·leccions temàtiques. El sistema de tractament efectua diferents processos que tenen com a finalitat validar la pertinença dels recursos capturats, i la posterior indexació i catalogació per tal de facilitar la recuperació d'informació per part dels usuaris. És també dins aquest sistema on es determina la freqüència de captura dels recursos que formen part del dipòsit, i que dóna lloc a un nou cicle d'entrada de documents que passen pels diferents processos que s'expliquen en els següents apartats:

4.2.1. Validació

La fase prèvia al tractament dels arxius haurà estat la selecció i captura semiautomatitzada de recursos, la planificació de la qual representa que haurà aportat un alt percentatge d'encert en la recopilació de webs en funció dels nostres criteris. Malgrat la confiança que es pugui tenir en la fiabilitat del sistema de selecció, els recursos hauran de passar per un filtre humà que s'encarregarà de validar la pertinença de cadascun dels webs en base a la temàtica del seu contingut. Per tal de prendre una decisió correcte sobre la validesa dels recursos i agilitzar-ne el seu procés, es crearà un patró dels criteris que haurà de complir un document per poder formar part d'una col·lecció temàtica concreta. Aquest patró de document haurà de definir aspectes com següents:

- Temes concrets: a part del tema principal de la col·lecció, s'hauran de definir temes més específics relacionats amb l'esdeveniment per tal de delimitar l'abast temàtic.
- Àrees geogràfiques: inicialment l'abast geogràfic sempre serà Catalunya, però en funció de l'esdeveniment a capturar es pot ampliar o reduir l'abast geogràfic dels recursos que seran validats.
- Períodes cronològics: bàsicament el període cronològic marcarà les dates extremes (inici i final) de la duració de la repercussió de l'esdeveniment. Els recursos amb una data d'actualització anterior a la data assenyalada com inici no seran acceptats perquè no contindran informació d'interès per la col·lecció. En l'altre extrem, encara que es continuï executant la captura, no es validaran els webs que hagin estat actualitzats després de la data que posa fi a la col·lecció.

- Personalitats relacionades amb l'esdeveniment: es crearan uns llistats de personatges vinculats directament amb l'esdeveniment que protagonitzarà la col·lecció per tal de validar els recursos que donin informació sobre aquests.
- Entitats relacionades amb l'esdeveniment: es crearan uns llistats d'entitats vinculades directament amb l'esdeveniment que protagonitzarà la col·lecció per tal de validar els recursos que donin informació sobre aquestes.
- Longitud dels continguts: degut a que valoració qualitativa dels webs té un alt component subjectiu i requereix molt temps, es farà una valoració quantitativa dels continguts dels recursos, havent de tenir un mínim de pàgines relaciones amb l'esdeveniment.
- Tipologia de documents prioritaris: bàsicament es recolliran formats de text, pdf, d'imatge i so. Els formats no estàndards tindran una prioritat molt baixa en el varem de captura.

Dels documents que no superin la fase de validació es generarà un llistat d'URLs per tal d'evitar el malbaratament de recursos, impedint que tornin a ser capturats en fases posteriors. En el cas dels documents que si superin la validació, passaran al següent estadi del tractament, la indexació.

4.2.2. Indexació

La indexació dels recursos serà un procés totalment automatitzat que es realitzarà amb un programa especialitzat en aquesta funció. La finalitat de la indexació és la creació d'un fitxer invers de termes significants de les pàgines web que permeti facilitar-ne la recuperació. A banda de la construcció del fitxer invers, en aquesta fase s'extrauran de manera automàtica dels recursos, mitjançant el mateix programari d'indexació, una sèrie de metadades que formaran part del registre de descripció. Les metadades extretes automàticament seran les següents:

- **Identificador**: es pren l'URI com identificador de cada recurs (Uniform Resource Identificator) ja que aquest és únic i irrepetible. L'URI no és un element concret sinó un conjunt lògic format pels elements URL (Uniform Resource Locator), URN (Uniform Resource Name) i URC (Uniform Resource Characteristic). Aquest identificador és utilitzat per la màquina per distingir cadascuna de les pàgines web que formen part del dipòsit i poder-les recuperar de forma aïllada.
- **Títol**: el títol s'extreu de la metaetiqueta <*titlte*> que figura en el <*head*> del codi font de cada pàgina del lloc web. Aquesta metadada, la qual no té perquè ser única

i irrepetible, té la finalitat d'identificar cadascuna de les pàgines amb un llenguatge intel·ligible per part de l'usuari.

- **Llengua**: aquesta metadada es dóna per defecte ja que en aquesta primera fase del projecte tots els recursos hauran de complir el criteri que estiguin escrits en català. El fet que la definició de Patrimoni Bibliogràfic Català inclogui també documents escrits en altres llengües diferents al català, obliga disposar de la metadada llengua per si aquests són capturats en properes fases del projecte.
- **Format de l'arxiu**: la metadada format d'arxiu (*mimetype*), a part de donar un valor descriptiu al registre del recurs, ha de servir per poder agrupar els documents en les futures accions de preservació, sobretot en el cas de les migracions.
- **Data de captura**: en el procés d'indexació es grava la data de captura del recurs. Aquesta metadada, a part de ser significativa per la gestió interna dels recursos, serà el punt d'accés pels usuaris finals de cadascuna de les versions captures que es realitzaran de cada lloc web.
- **Resum**: davant la impossibilitat que els agents editors del projecte realitzin un resum de cada lloc web, donat al gran nombre de recursos que s'hauran de processar, es determina que s'extregui de forma automàtica un fragment de text inicial del web, per tal de tenir una mostra del seu contingut.

4.2.3. Catalogació

La catalogació dels webs la realitzarà un agent humà especialitzat en tasques de descripció de recursos electrònics. Aquest agent interactuarà amb una interfície d'edició que facilitarà l'agregació de les metadades en una base de dades que les relacionarà amb el recurs digital corresponent.

Seguint l'exemple del projecte PADICAT i altres projectes relacionats, es farà servir el patró de metadades del Dublin Core per descriure el corpus de recursos que formaran part del dipòsit del TematiCAT. Tenint en compte la quantitat de llocs web que s'hauran de catalogar i les metadades necessàries per poder-ne assegurar una correcte recuperació, s'ha optat per seguir el nivell mínim de descripció que aconsella Dublin Core, el qual està format per les metadades: títol, data creació, data de captura, llengua, matèria i resum.

Exceptuant el cas de la matèria, la resta de metadades s'extrauran de forma automàtica del web en qüestió.

La metadada matèria es generarà usant la Llista d'encapçalaments de matèria en català (LEMAC)[83], que es farà servir com a vocabulari controlat. Tot i així, no s'emprarà la LEMAC de forma canònica, ja que no es pretén generar concatenacions de matèries amb subdivisions. Únicament s'utilitzaran els termes *acceptats*, que actuaran com a descriptors, i no es tindran en compte els termes que són *subdivisions* i *no acceptats*. També s'introduiran noms de personalitats i entitats com a descriptors, que seran validats mitjançant la Llista d'encapçalaments de noms i títols (LENOTI)[84]. Ambdues llistes es podran ampliar amb nous termes i noms, que seran validats a priori.

Aprofitant que les col·leccions que constituiran aquest projecte són temàtiques, prèviament a la descripció dels llocs web, es crearà un llistat de descriptors possibles per a cada col·lecció. D'aquesta manera s'agilitzarà el procés de catalogació, i es propiciarà que existeixi un nivell alt de coherència amb els termes seleccionats. D'altra banda, la limitació del nombre de descriptors possibles permet fer servir el camp de matèria per crear clústers d'informació, podent elaborar, per exemple, subcategories temàtiques de webs.

Una de les principals dificultats a l'hora de definir la descripció de la metadada matèria ha estat entendre si s'havien de catalogar les pàgines web o els llocs web. Inicialment ja queda clar que la catalogació de les pàgines de forma individual és una empresa faraònica, però el fet de poder trobar webs que només tinguin un interès parcial, o dit d'altra manera, que només siguin pertinents algunes de les seves pàgines, ens porta a pensar si aquestes han de formar part de les col·leccions, i en el cas que s'acceptin, com s'han de catalogar.

Prenen com exemple els esquemes clàssics de la catalogació de documents físics, entenem que la prioritat és la unitat de l'entitat (web) i llavors la recuperació d'informació. Usant un exemple real, un manual de marketing, que té com a matèria *marketing* i *manuals, guies, etc.*, podria tenir també el descriptor *segmentació*, perquè hi ha alguns capítols del llibre que estan dedicats a aquest aspecte concret. En el procés de la catalogació dels webs es pot aplicar el mateix patró, ja que el que interessa és recuperar l'entitat que conté la informació, i llavors de la mateixa manera que un llibre té un sumari, la web té eines per poder cercar els continguts. A patir

[83] BIBLIOTECA DE CATALUNYA. *Llista d'encapçalaments de matèria en Català* [En línia]: LEMAC. Barcelona: Biblioteca de Catalunya, 2005. [Data de consulta: 02/05/2007]. Disponible a: <http://www.bnc.es/catalegs/autoritats/lemac.php>.

[84] BIBLIOTECA DE CATALUNYA. *Llista d'encapçalaments de noms i títols* [En línia]: LENOTI. Barcelona: Biblioteca de Catalunya, 2005. [Data de consulta: 02/05/2007]. Disponible a: <http://www.bnc.es/catalegs/autoritats/lemac.php>.

d'aquest raonament s'extreu que quan es descriu la metadada de matèria aquesta afecta tot el lloc web, i que quan es recupera una pàgina web en concret, aquesta hereta tots els descriptors que se li han donat al web sencer, entenent que aquella pàgina està relacionada amb un tot.

Pel que fa a la metadada de autor o contribuïdor, després d'analitzar detingudament la problemàtica de com descriure-la[85], s'ha optat per no incloure-la en el registre catalogràfic. Molts dels llocs web que seran objectiu de captura per a les col·leccions del projecte TematiCAT són fruit de col·laboracions de múltiples autors, que algunes vegades són agrupats sota entitats, però en moltes altres no. Davant la impossibilitat de mantenir una coherència en la descripció del camp d'autor, i per la poca rellevància que té aquesta dada en la recuperació d'informació en la xarxa, en la fase de planificació del projecte no se'n contempla la descripció.

4.2.4. Programació de la freqüència de captura d'un lloc web

L'última tasca del tractament d'un web, abans de ser allotjat definitivament al dipòsit del TematiCAT, és la determinació de la freqüència de captura d'aquest, entenen que serà actualitzat periòdicament. Aquesta és una tasca subjectiva, i per tant realitzada novament per un agent humà, que mitjançant un software pot programar el *crawler* perquè visiti i capturi en les dates establertes un recurs en concret.

De la mateixa manera que en processos anteriors, es crearà un patró de freqüències de captura dels recursos, per tal de minimitzar els efectes del subjectivisme. Aquest patró donarà unes pautes de programació en funció de la tipologia del recurs, ja que, per exemple, el web d'un mitjà de comunicació requerirà una freqüència de captura més alta (d'una o més vegades per dia), mentre que un web relativament estàtic serà suficient que es visiti un cop a la setmana mentre duri el procés de recopilació.

Una vegada establerts els períodes de captura en la consola d'administració dels recursos, aquesta executa de forma automàtica el robot que recollirà els arxius provinents d'Internet, i els allotjarà temporalment en l'espai de treball virtual. S'ha de matisar que en aquesta fase de recopilació de versions, el robot que realitza la captura estarà programat per verificar que els recursos han estat actualitzats, ja que en cas contrari, si es tracta de fidels duplicats dels anteriors, aquests no seran capturats. Els recursos que superin el filtre de control de duplicats passaran a la fase de tractament, on seran validats i indexats. Pel que fa a la descripció dels recursos, aquests rebran

[85] INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *ISBD (ER)* [En línia]: International Standard Bibliographic Description for Electronic Resources. França: IFLANET, juliol de 1999 [Data de consulta 06/06/2007]. Disponible a: <http://www.ifla.org/VII/s13/pubs/isbd.htm>.

per defecte els descriptors que s'hauran donat al web original, amb la possibilitat que l'agent editor pugui realitzar les pertinents modificacions si és necessari.

4.2.5. Requeriments tecnològics del programari de tractament de recursos del dipòsit PADICAT

El programa encarregat del tractament dels recursos capturats haurà de complir els següents requeriments per tal de poder ser adaptat a la plataforma tecnològica del projecte TematiCAT:

A nivell d'implementació

- Programari de codi obert per tal de ser desenvolupat posteriorment en cas que sigui necessari. Els òrgans de l'administració pública recomanen fer extensiu l'ús de programari lliure
- Compatibilitat amb el sistema operatiu Linux. Com en el cas del programa que es vol adquirir, el sistema operatiu que suportarà tota la plataforma tecnològica serà de codi obert.
- Compatibilitat amb el servidor d'aplicacions web Tomcat 5.0
- Compatibilitat amb bases de dades MySQL
- Compatibilitat amb aplicacions de recuperació i visualització de recursos web

A nivell de funcionalitats

- Capacitat d'executar i parametritzar treballs de captura de recursos web
- Capacitat d'establir freqüències de captura automatitzades de determinats recursos web
- Capacitat d'agregar metadades als recursos web
- Capacitat de gestionar arxius en format comprimit

4.2.6. Estudi del programa d'administració d'arxius Web Curator

Web Curator és un programa que participa en els processos de recopilació d'arxius web i permet la posterior administració d'aquests. Una de les seves principals característiques és la interfície d'administració, amb un disseny amigable i un sistema de gestió usable, que permet el tractament dels recursos sense requerir als agents editors un alt nivell de coneixements tecnològics de l'entorn web. Web Curator es

concep com un sistema de workflow (*Annex 10*) que assisteix des de la selecció dels webs a capturar, passant pel tractament d'aquest, fins arribar al seu allotjament permanent en el dipòsit.

Web Curator va ser desenvolupat, i presentat el setembre de 2006, de la mà de la National Library of New Zeland i la British Library, amb la col·laboració de l'empresa Sytec ,que va realitzar el desenvolupament tecnològic del sistema.

Aquesta aplicació, com la resta de programes que s'han presentat en aquest projecte, ha estat desenvolupada en codi obert sobre el llenguatge de programació Java. És compatible amb la majoria de sistemes operatius, entre ells Linux, i funciona sobre la plataforma Apache i Tomcat. En el disseny de la multiplicitat de funcions que realitza Web Curator, s'han integrat alguns programes ja existents, o part d'ells, que efectuen algunes tasques concretes[86]. Entre les diferents funcionalitats de Web Curator destaquem les següents[87]:

- **Autorització de captura**: aquesta funció està pensada pels projectes que contacten amb els editors dels webs objectiu per sol·licitar permís de captura abans de procedir a la recopilació. Per portar a terme aquesta funció es creen uns registres on figuren les pàgines que es volen capturar, juntament amb les dades referents a la organització o autor que publica el domini. Una vegada rebuda la resposta el sistema permet executar les captures en funció de la política de permisos.
- **Creació de *targets***: una vegada s'ha rebut l'autorització de l'autor del recurs, en el cas que aquest es sol·liciti, l'aplicació permet programar treballs (*jobs*) que defineixen exactament què i quan s'ha de capturar. Es genera una fitxa per cada recurs on figura el nom, l'organització o autor que el publica, una breu descripció i el seu estat dins el cicle de captura. Les dades d'aquests registres permeten la parametrització de treballs de forma individual o d'un conjunt de *targets*. Una de les possibilitats de configuració més rellevant és la freqüència de captura (*scheduling*) dels *targets*, podent especificar a partir de quan s'ha d'iniciar la recopilació i la periodicitat amb que es farà. En tot moment Web Curator permet la revisió dels

[86] El programes que s'han incorporat a Web Curator són aplicacions que han estat dissenyades per realitzar determinades tasques de forma individual. Entre les aplicacions agregades hi ha programes ja coneguts com Heritrix o Way Back, entre d'altres com Acegi Secutity System, Apache Axis (SOAP data transfer), Apache Commons Logging, Hibernate (database connectivity), Quartz (scheduling) i Spring Application Framework.

[87] NATIONAL LIBRARY OF NEW ZEALAND; BRITISH LIBRARY. *Web Curator Tool Quick Start Guide*[En línia]. National Library of New Zealand & British Library, setembre de 2006. [Data de consulta: 03/04/2007]. Disponible a: <http://webcurator.sourceforge.net/docs/1.1/wct-1.1-quick-start-guide.pdf>.

processos de captura i la modificació dels paràmetres de configuració si és necessari.

- **Descripció dels webs**: tot i no tractar-se d'un sistema de catalogació, Web Curator disposa d'una funcionalitat que permet agregar metadades de descripció Dublin Core als objectes capturats.

Web Curator no té la capacitat de recuperar recursos del dipòsit i visualitzar-los a través de la xarxa, però tanmateix ofereix plena compatibilitat amb programes que realitzen aquestes tasques com Wera o Wayback.

4.3. Disseny del sistema de recuperació i visualització dels recursos del dipòsit TematiCAT

El sistema de recuperació i visualització estarà format per una sèrie d'aplicacions que tenen com a finalitat permetre a l'usuari poder accedir als recursos que formen les col·leccions del TematiCAT. Aquesta recuperació d'informació per part de l'usuari serà possible gràcies a la seva interacció amb el sistema per mitjà d'una interfície, amb la qual podrà realitzar cerques al dipòsit de recursos i visualitzar els resultats obtinguts.

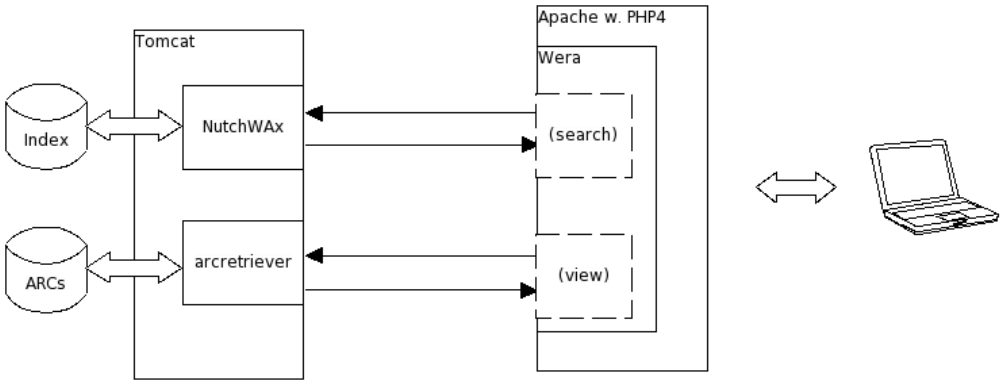


Fig. 22. Arquitectura de processos del sistema de recuperació i visualització

L'aplicació que actuarà com a motor de cerca és la mateixa que s'utilitzarà en la preselecció de recursos del dipòsit del PADICAT durant la fase de captura. Aquest programa de cerca permetrà interrogar els diferents camps de descripció, als quals s'hauran agregat les metadades en la fase de tractament, o bé fer cerques per mitjà de paraules clau en base el fitxer invers creat en el procés d'indexació.

Els ítems resultants de la cerca seran transmesos a l'aplicació de visualització, la qual generarà un llistat de registres de recursos, on hi podran figurar dades com el títol, la URL i un fragment de text de la pàgina web. Aquests registres donaran accés a recursos concrets allotjats en el dipòsit, els quals podran ser navegats sempre i quan s'hagin capturat les pàgines a les que apunten els enllaços. A part de donar accés als webs obtinguts en l'operació de cerca, l'aplicació de visualització permetrà poder accedir a totes les captures d'un recurs en concret.

4.3.1. Requeriments tecnològics del programari de visualització de recursos del dipòsit PADICAT

Tal com ja s'ha comentat, el programa que actuarà com a motor de cerca serà el mateix que es farà servir en la fase de captura. Per altra banda, hi ha dues possibilitats d'aplicacions que es poden implementar en el mòdul de visualització, Wera i Wayback machine, de les quals tot seguit es detallen els requisits.

A nivell d'implementació

- Programari de codi obert per tal de ser desenvolupat posteriorment en cas que sigui necessari. Els òrgans de l'administració pública recomanen fer extensiu l'ús de programari lliure
- Compatibilitat amb el sistema operatiu Linux: com en el cas del programa que es vol adquirir, el sistema operatiu que suportarà tota la plataforma tecnològica serà de codi obert.
- Compatibilitat amb el servidor d'aplicacions web Tomcat 5.0
- Compatibilitat amb el servidor de pàgines web Apache 2.0
- Compatibilitat amb bases de dades MySQL
- Compatibilitat amb determinades aplicacions de cerca de recursos web

A nivell de funcionalitats

- Capacitat de generar llistats de resultats de cerques de recursos web
- Capacitat de visualització i navegació de recursos webs
- Capacitat d'oferir llistats de les diferents captures d'un web

4.3.2. Estudi dels programes de recuperació i visualització

Wera

Wera és una aplicació de visualització preparada per donar accés als recursos de les col·leccions d'un dipòsit digital, alhora de permetre una fàcil navegació entre les diferents versions dels webs[88]. Wera va ser desenvolupat pel Nordic Web Archive (NWA), amb l'esponsorització de la IIPC, en substitució de l'antic programari NWA Toolset, i va ser presentada la seva primera versió l'agost de 2005. Wera es va desenvolupar usant els llenguatges PHP i Java, i utilitza *open standards* com el protocol http i XML per comunicar els diferents mòduls del sistema.

Per tal de recuperar i fer accessibles els recursos als usuaris, Wera treballa conjuntament amb dos tipus d'aplicacions: un motor de cerca, que inicialment va ser FAST, i que actualment és Nutch Wax; i un agent de recuperació d'informació (d*ocument retreiver*), que s'encarrega de mostrar els recursos arxivats juntament amb les seves metadades en la interfície d'usuari.

Wera posa a la disposició de l'usuari diferents estratègies d'interrogació, a través de la cerca simple i la cerca avançada, que estan dirigides als diferents camps de descripció dels recursos arxivats. Aquesta *query* es trasllada al motor de cerca Nutch Wax, que l'executa en la base de dades on s'allotgen les metadades i els termes del fitxer invers. La resposta amb els resultats, en format XML, és retornada a Wera que transforma el llenguatge rebut en un format propi en PHP. Finalment els arxius són recuperats a través del *document retriever*, i són visualitzats en la interfície, on l'usuari pot determinar l'ordre d'aparició. La pantalla de la llista de resultats permet navegar els webs de diverses formes a través de les opcions *view*, *timeline* i *more from this site*. La primera d'aquestes opcions obre una pàgina web obtinguda en la cerca; la segona, *timeline*, ofereix un llistat de les captures disponibles d'aquell lloc web en concret; finalment, *more from this site*, mostra totes les pàgines web arxivades associades a un lloc web en particular. Quan es visualitza un recurs aquest es mostra dins un *framework* de Wera, on es plasmen determinades dades com al URL o la data de captura de l'arxiu, i alhora dóna la possibilitat de navegar la resta de versions recopilades mitjançant uns enllaços incrustats en una línia de temps.

[88] INTERNACIONAL INTERNET PRESERVATION CONSORTIUM; NORDIC WEB ARCHIVE. *Wera* [En línia]. Internet Archive, 2005. [Data de consulta 04/05/2007]. Disponible a: <http://archive-access.sourceforge.net/projects/wera/>.

Wayback Machine

Wayback machine és el programa de visualització de recursos web arxivats usat per Internet Archive[89]. La primera versió d'aquesta aplicació va veure la llum l'any 2001 amb la finalitat de fer accessibles tots els recursos que havien estat capturats des de l'any 1996. Wayback machine es caracteritza per tenir una interfície molt senzilla i austera, però que malgrat això, el disseny d'aquesta ha estat la font d'inspiració per la majoria de programes que han realitzat funcions similars.

En el procés de cerca dels recursos, Wayback Machine únicament accepta les URLs com a punt d'accés als recursos arxivats, i no permet la cerca a través de paraules clau. Tanmateix, si que es contempla la possibilitat de realitzar una cerca simple, usant una adreça exacte, o una cerca avançada. En aquesta segona opció també s'ha de buscar en base una URL, però en aquest cas ens permet realitzar certes operacions amb ella, com la recuperació de recursos que no tinguin exactament aquella URL, mostrar tots els recursos arxivats d'una domini, administrar les pàgines que redirigeixen a una altra, ometre pàgines duplicades, o bé fer cerques segons el criteri de format. Pel altra banda, l'assistent de cerca disposa d'uns camps de limitació cronològica per poder determinar el període de captura del web en qüestió.

El resultat d'una cerca es tradueix en forma de calendari on es mostren les dates de les captures, les quals són l'enllaç perquè es visualitzi la pàgina web.

4.3.3. Prototip de la interfície de cerca l'usuari

La interfície serà el pont entre l'usuari i el sistema de recuperació i visualització. Tal com s'ha comentat en la introducció, el disseny de la interfície, contemplant aspectes com la usabilitat, l'accessibilitat i el disseny gràfic, no forma part de l'abast del projecte. Tot i així, en aquest apartat es definiran, mitjançant un prototip esquemàtic, quines seran les eines de cerca que permetran recuperar els recursos del dipòsit i quins altres tipus de punts d'accés s'oferiran a la informació, degut a que el seu funcionament està estretament relacionat amb la configuració del sistema d'informació.

Pel que fa a les eines de cerca es preveu que es disposi d'un apartat de cerca simple, en la qual només es podrà buscar a través de paraula clau, i un apartat de cerca avançada que permetrà interrogar els diferents camps de descripció dels recursos. La cerca avançada contindrà els següents camps de consulta:

[89] NOTESS, GREG R. "The Wayback Machine [En línia]: the Web's Archive". *Info Today*, vol 26, núm. 2, març-abril de 2002. [Data de consulta: 04/05/2007]. Disponible a: <http://www.infotoday.com/online/mar02/OnTheNet.htm>.

- **Col·lecció:** abans d'introduir la interrogació en els camps de consulta, l'usuari podrà escollir en quina col·lecció vol cercar, o en tot cas, si vol cerca en totes alhora. Per defecte apareixerà l'última col·lecció creada.
- **Paraula clau:** aquest camp serà de cerca a text lliure
- **Adreça del web (URL):** l'usuari podrà recuperar un web concret introduint la seva adreça.
- **Matèria:** aquest camp serà un desplegable on apareixeran les matèries referents a la col·lecció seleccionada (o de totes les col·leccions en cas que s'hagi seleccionat aquesta opció), on l'usuari les podrà escollir per formular la consulta. La limitació d'aquest camp és que només es podran consultar les matèries d'una en una, ja que la tecnologia del camp no en permet seleccionar més d'una alhora.
- **Dates:** aquest camp permetrà delimitar cronològicament l'abast de la consulta mitjançant la introducció de dos dates extremes que definiran el període dels webs que seran retornats.

Tant en la cerca senzilla com en la cerca avançada es permetrà l'ús d'operadors booleans (*and*, *or* i *not*) i de truncament (*[asterisc]*). L'operador *or* actuarà per defecte entre els termes en les cerques a text lliure per facilitar la consulta als usuaris. En el cas que no hagi cap recurs que respongui a la consulta efectuada apareixerà un quadre de diàleg avisant que la cerca no ha produït resultats, informant també que el dipòsit conté una selecció de recursos de la xarxa i que existeix la possibilitat que el web que cerca l'usuari no hagi estat recopilat. Finalment hi haurà un vincle cap a un formulari de la interfície on es permet suggerir recursos.

A part de les eines de cerca que es proporcionaran a l'usuari, s'ubicaran en la interfície quatre mòduls destinats a oferir diferents punts d'accés a la informació referent a la última col·lecció creada, en base diversos criteris que podrà gestionar l'administrador del web TematiCAT. Aquests mòduls tindran la funció de donar accés als recursos des d'altres perspectives diferents la cerca, però sobretot jugaran un paper molt important en fer més dinàmica la interfície, ja que periòdicament s'hauran d'actualitzar amb nous continguts. Alguns exemples de la informació que poden contenir aquest mòduls són els següents:

- **Webs més visitats:** aquest mòdul contindrà un llistat dels llocs web més visitats pels usuaris. Aquests llistats s'obtindran a partir de l'anàlisi dels *logs* de la interfície del TematiCAT, on es podrà observar quins han estat els recursos més visitats. A partir de les paraules clau introduïdes o matèries seleccionades en les cerques, es podrà veure quines han estat les consultes més freqüents dels visitants.

- **Selecció de personalitats relacionades amb l'esdeveniment**: aquest mòdul recollirà recursos on es mostrarà informació referent a personatges directament relacionats amb els esdeveniments. Per oferir aquesta informació, l'administrador de la interfície del TematiCAT realitzarà una cerca en el camp de matèria, en el qual s'hauran introduït els noms dels personatges.
- **Selecció d'entitats relacionades amb l'esdeveniment**: de la mateixa manera que s'ofereix un llistat de personalitats relacionades amb l'esdeveniment, es podrà oferir una relació de recursos amb informació de les entitats directament vinculades. Aquesta llista de recursos també s'obtindrà del camp matèria on hauran estat introduïts els noms de les entitats.
- **Webs proposades**: per tal de fer patent la col·laboració dels usuaris amb el projecte TematiCAT, es crearà un mòdul destinat a reflectir els webs suggerits per usuaris.
- **Altres col·leccions del TematiCAT**: aquest mòdul servirà per mostrar les col·leccions de recursos relacionats amb esdeveniments passats, per tal que tinguin una contínua visibilitat.

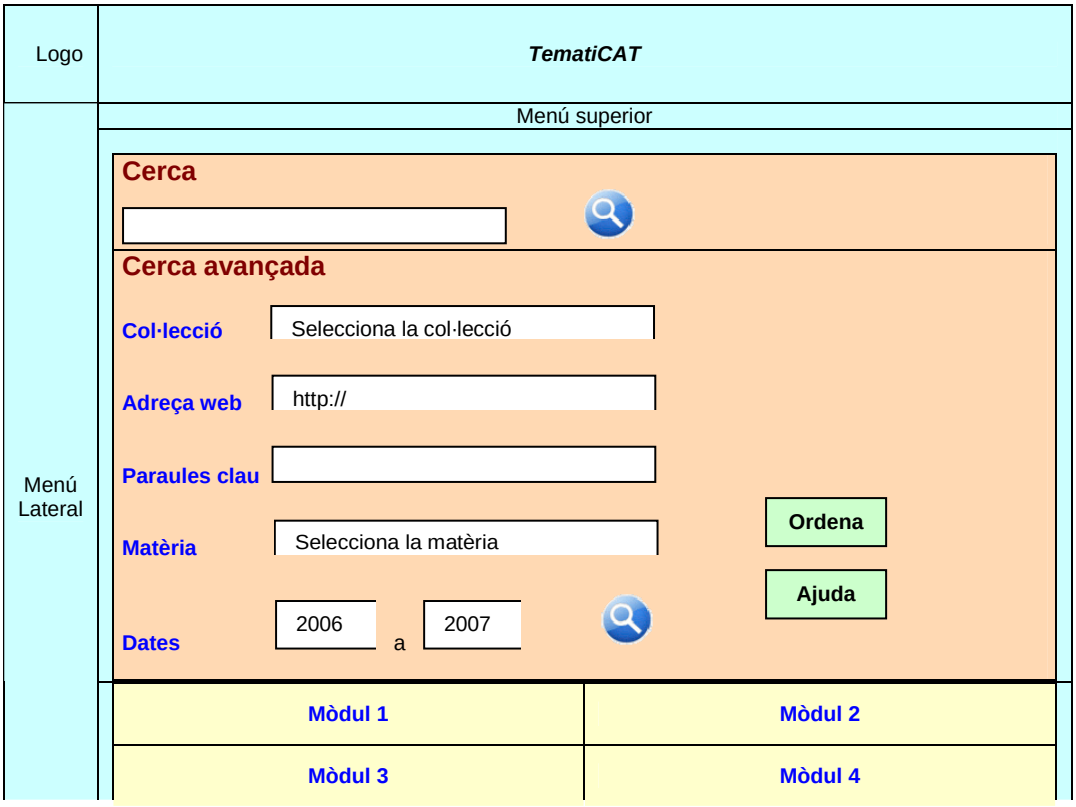


Fig. 23. Prototip de la interfície de cerca del TematiCAT

En la imatge superior es mostra un prototip esquemàtic de la interfície de cerca de l''usuari del TematiCAT amb la finalitat de plasmar els diferents punts d'accés als recursos. La part ombrejada en color blau són els components de la interfície que no són tractats en aquest projecte, i que hauran de ser desenvolupats per un equip especialitzat en disseny gràfic, usabilitat i accessibilitat. La zona ombrejada en color taronja serà l'apartat destinat a la cerca de recursos, el qual es dividirà en la cerca simple i la cerca avançada. En aquest espai hi haurà un opció d'ordenació dels resultats de la cerca i també s'hi podrà consultar una secció d'ajuda en la cerca. Finalment, la part ombrejada en color groc serà l'espai destinat a contenir els diferents mòduls que donaran accés als recursos a partir de les possibilitats que s'han citat anteriorment.

5. Implementació del sistema d'informació

La implementació de l'arquitectura del sistema d'informació es realitzarà en base els models estudiats d'altres projectes i l'anàlisi dels diferents programes que compleixen els requeriments de cada mòdul. A l'hora de prendre la decisió de quins són els programes que composaran el sistema s'ha tingut en compte la relació amb els requeriments establerts i la compatibilitat amb el sistema del PADICAT, amb el qual haurà d'interactuar en determinats processos. També s'ha tingut en compte el grau de compatibilitat que existeix entre els diferents programes que formaran el sistema, per tal d'evitar problemes de comunicació entre ells, cosa que podria significar un increment considerable del pressupost en desenvolupament tecnològic.

La cadena de processos del sistema, dividida en els mòduls d'ingestió de recursos, tractament i accés a les col·leccions del dipòsit, estarà suportada pels programes Heritrix (captura), Nutch Wax (indexació i consulta), Web Curator (tractament) i Wera (recuperació i visualització), quedant l'arquitectura del sistema definida de la següent manera:

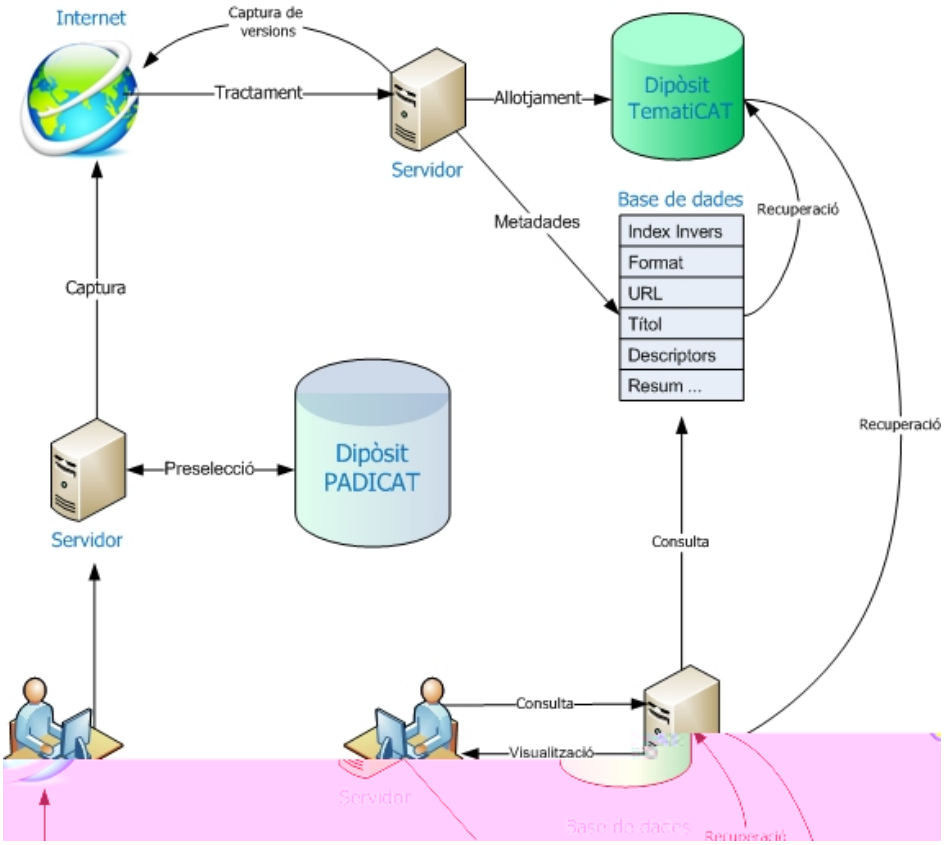


Fig. 24. Esquema del sistema d'informació del TematiCAT

En els següents apartats d'aquest capítol s'explicarà quin serà el procediment d'implementació de les aplicacions, i quins desenvolupaments tecnològics i proves pilot s'hauran de realitzar per adaptar cada programa a les necessitats del projecte. En el cas de les proves pilot s'explica quines són les accions a realitzar per comprovar el correcte funcionament de les aplicacions, però no es pretén detallar el nombre de proves i l'exactitud de com es farà, ja que en funció de cadascun dels tests es determinarà continuar, canviar d'estratègia o donar per bona la implementació. Per aquest mateix motiu es dóna un *timing* global per cadascun dels tres sistemes, ja que resulta molt difícil poder afirmar amb seguretat la duració de la implementació i proves que es realitzaran. Les estimacions previstes de la duració de la implementació contempla que algunes de les tasques es poden allargar, i de la mateixa manera, altres poden ser executades en menys temps de l'indicat.

5.1. Implementació del sistema operatiu i el programari base

Abans d'iniciar la implementació dels programes funcionals del sistema d'informació, es realitzarà la instal·lació de la plataforma del programari base (sistema operatiu i aplicacions de servidor). Aquest projecte aposta clarament per les aplicacions de codi obert, per aquest motiu el sistema operatiu que controlarà tot el sistema serà Linux, Apache 2 serà el programa vinculat amb el servidor web i Tomcat 5 amb el servidor d'aplicacions.

5.2. Implementació del mòdul d'ingestió de recursos web

El mòdul d'entrada de recursos web realitzarà les funcions de preselecció de recursos del dipòsit PADICAT, en base a una consulta prèvia, i la captura de webs provinents d'Internet a partir de diferents estratègies de cerca. Els programes escollits per portar a terme les funcions d'aquest mòdul són Heritrix i Nutch Wax.

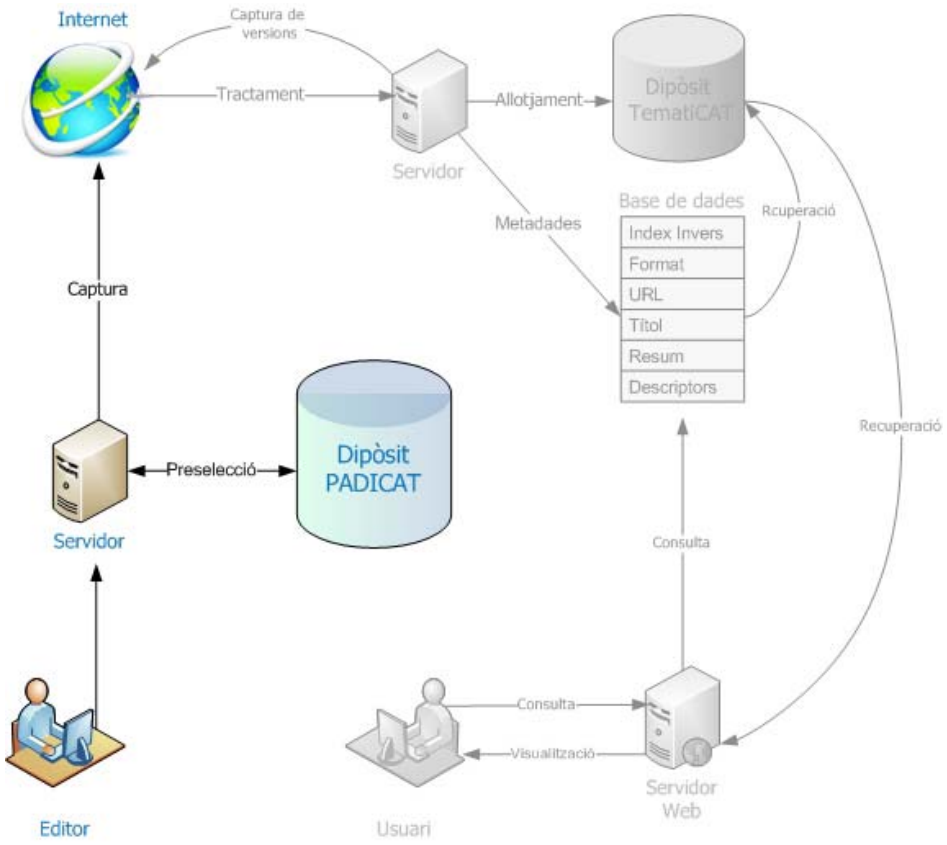


Fig. 25. Detall del sistema de selecció i captura dins el sistema global

En la selecció del programa encarregat de capturar els webs d'Internet, Combine i Heritrix han estat les dos aplicacions que han resultat més interessants per les característiques del projecte, i de les quals s'ha optat finalment per la segona. Combine ha estat implementat en diferents projectes, i ofereix unes prestacions molt avançades en el camp de la captura focalitzada de recursos, no obstant, es desconeix el grau de compatibilitat amb la plataforma tecnològica del PADICAT, i amb la resta de programes integrants dels sistema d'informació del TematiCAT. Per altra banda, Heritrix és el mateix programa que usa el mòdul de captura del PADICAT i és completament compatible amb les aplicacions Nutch Wax i Web Curator, amb les quals haurà d'interactuar.

Nutch Wax serà el programa amb el que es realitzarà la preselecció de recursos del dipòsit PADICAT, a través del seu mòdul de consulta. Com en el cas del programa Heritrix, Nutch Wax és una aplicació implementada en diversos dels projectes que s'han estudiat i rep l'emparament de l'IIPC, a més de ser uns dels programes usats pel PADICAT.

El funcionament del mòdul d'ingestió de recursos web, a partir de la implementació dels programes Heritrix i Nutch Wax, queda definit de la següent forma:

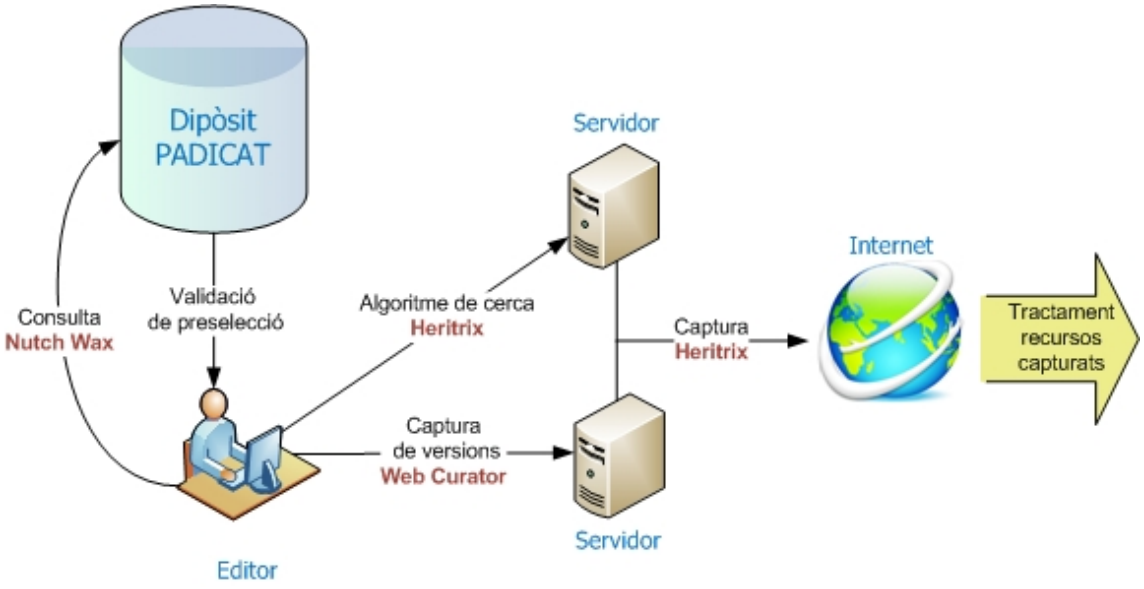


Fig. 26. Sistema de selecció i captura

5.2.1. Implementació i adaptació de Nutch Wax

La implementació de l'aplicació Nutch Wax té prevista una duració de 70 hores aproximadament. En aquesta fase d'instal·lació del programa, dins el *framework* del mòdul d'ingestió de recursos, es parametritzaran les següents funcions:

- Consulta de recursos del dipòsit PADICAT: la implementació de l'aplicació ha d'assegurar que es podran realitzar cerques en determinats camps per tal de realitzar la preselecció. Els camps que seran objecte d'interrogació són els de matèria, títol i fitxer invers, amb els que s'identificaran els recursos potencialment interessants, i el d'URL, amb la qual es podrà recuperar el recurs. En l'adaptació de la funció de consulta al dipòsit s'haurà de comprovar que es poden realitzar cerques complexes, mitjançant la utilització d'operadors booleans i de truncament, i alhora la possibilitat de limitar cronològicament a partir del camp de data de captura dels arxius.
- Capacitat de lectura d'arxius en format comprimit: els recursos allotjats en el dipòsit PADICAT estan comprimits, i per tant s'haurà d'integrar a Nutch Wax una aplicació que li permeti administrar aquest tipus de formats. BAT[90] (BnF Arc Tools), és una aplicació que permet administrar i modificar arxius en format ARC, DAT i CDX. Mitjançant les instruccions de BAT, el motor de cerca de Nutch Wax podrà realitzar

[90] Hafri, Younès. BAT [En línia] BnfArc Tools. Paris: Bibliothèque Nationale de France, octubre de 2005. [Data de consulta: 01/02/2007]. Disponible a: <http://bibnum.bnf.fr/downloads/bat/>.

consultes als recursos en format comprimit del dipòsit sense la necessitat d'haver-los de descomprimir per accedir al seu contingut.

Per tal de confirmar la correcta configuració del sistema de cerca de Nutch Wax, es realitzaran una sèrie de proves pilot en les que s'analitzarà el comportament del motor de cerca a l'hora d'efectuar cerques en el dipòsit PADICAT. En aquests assajos es comprovarà l'eficiència de cerca mitjançant els operadors booleans i de truncament, així com la possibilitat d'establir límits cronològics mitjançant el camp de data de captura.

5.2.2. Implementació i adaptació d'Heritrix

La implementació de l'aplicació Heritrix té prevista una duració aproximada de 70 hores. Heritrix és un programa caracteritzat per la seva flexibilitat i la diversitat de les seves funcions. Per tal d'aprofitar les prestacions que aporta aquest programa és molt important realitzar una acurada configuració dels seus paràmetres de funcionament. En aquesta fase s'instal·larà el programa i es comprovarà que les següents funcionalitats responen a les necessitats del projecte En totes les proves pilots que es realitzin es prendran estadístiques de la duració dels processos per tal de fer estimacions de la duració dels treballs de captura.

- Captura de recursos per mitjà d'URLs: aquesta funció està destinada a capturar un o més recursos, en funció del nombre d'adreces que li proporcionem al programa. S'haurà de comprovar que s'efectua correctament la descàrrega d'un recurs, del qual li haurem proporcionat l'adreça, i què succeeix quan se li dóna una adreça que ja no està activa o ha estat redirigida. És important saber quines poden ser les conseqüències i com reaccionarà el *crawler* al donar una coordenada falsa, ja que les direccions canvien contínuament.

- Captura de recursos per mitjà d'IPs: aquesta funció és interessant quan no només volem capturar un lloc web, sinó que interessa tot el que s'allotja en una determinada direcció IP. No serà l'estratègia habitual de captura, però s'ha de comprovar que aquesta funcionalitat és viable.

- Identificació i captura de recursos a través de filtres: aquesta funcionalitat es basa en la captura de determinats recursos en funció del compliment d'una condició. En aquest cas s'hauran de configurar els paràmetres dels mòduls interns d'Heritrix per ordenar-li que cerqui un sèrie de termes en llocs concrets del web (títols,

encapçalaments, etc.). S'haurà de comprovar el nivell de silenci o soroll informatiu que retorna en funció dels graus de restricció dels filtres. Està previst que es realitzin diverses proves pilot en base a diferents àmbits temàtics per obtenir dades comparatives per tal d'establir el grau d'èxit d'aquesta estratègia de cerca, ja que és la més equilibrada pel que fa cost i retorn d'informació. Aquesta requerirà que es realitzi un mínim desenvolupament de programació per tal d'adaptar l'estratègia de cerca a les possibilitats del programa.

- Identificació i captura de recursos a través d'algoritmes: aquesta funcionalitat no és pròpia del programa, sinó que haurà de ser desenvolupada per un especialista en llenguatges de programació i un especialista en plantejament de models matemàtics. La cerca a través d'algoritmes vectorials només es plantejarà en el cas que es demostri que la resta d'estratègies no són valides. Tal com es va explicar en el seu apartat corresponent, la cerca per algoritme vectorial parteix del càlcul de freqüència d'aparició i la distància entre uns determinats termes. El model que s'obté d'aquest càlcul representa el patró que s'executarà a través del *crawler*. El problema d'aquesta estratègia respon a que per cada col·lecció temàtica s'haurà de desenvolupar un algoritme. Per aquest motiu, la prova pilot només es realitzarà en un sol àmbit temàtic, ja que la finalitat d'aquesta és comprovar que l'aplicació és capaç d'executar un cerca a través d'algoritmes vectorials.

- Profunditat de captura: s'haurà de configurar i posar a prova l'opció de limitar la profunditat de captura dels recursos web. Inicialment, per les característiques del projecte, interessa que el *crawler* descarregui totes les pàgines d'un lloc web, però tanmateix, a menys que se li ordeni el contrari, no capturi més enllà dels enllaços externs incrustats en els documents.

- Expansió viral: de la mateixa manera que Heritrix permet limitar la profunditat de captura d'un web, permet alhora ordenar al robot que segueixi navegant a partir dels enllaços que troba en les pàgines. Aquesta funcionalitat s'utilitzarà combinada amb alguna de les estratègies de cerca, per tal d'identificar nous recursos que no estaven a la llista d'URLs inicials.

- Regulació de la descàrrega d'arxius: en els apartats de l'anàlisi de projectes hem pogut observar que es té molt en compte la seguretat dels servidors que allotgen els recursos objectius. El sistema d'ingestió de recursos del TematiCAT pretén oferir les mateixes garanties, però tanmateix, els paràmetres de seguretat no tenen

perquè ser els mateixos que els observats, ja que la infraestructura tecnològica pot ser diferent. Per aquest motiu, mentre es realitzen assajos de funcionament del sistema, es prendran estadístiques sobre el comportament del *crawler* envers els servidors. En funció de les dades obtingudes, es limitarà el nombre de recursos i el temps de descàrrega.

- Compressió dels arxius capturats: s'haurà de comprovar que la compressió dels arxius capturats és òptima i compatible amb l'administrador BAT.

- Captures periòdiques automàtiques: s'haurà de comprovar que hi ha una correcte comunicació entre les aplicacions Heritrix i Web Curator, ambdues relacionades amb aquesta tasca. Es detallaran les proves pilot en l'apartat corresponent a l'aplicació Web Curator, dins el marc del sistema de tractament de recursos.

- Control de duplicats: en relació a les captures periòdiques de recursos, dins l'àmbit de treball de la fase de tractament, s'haurà d'implementar al programa Heritrix el mòdul DeDuplicator, que ha de permetre poder realitzar un control de duplicats dels webs que tenen programada una freqüència de captura. Aquesta aplicació s'ha d'adaptar a la base del programa Heritrix, i a l'hora s'ha de configurar i coordinar amb la base de dades on es troben els fitxers inversos de cada recurs (creats per Nutch), els quals seran la base de la comparació entre els webs.

5.3. Implementació del mòdul de tractament de recursos web

En el mòdul de tractament de recursos web es realitzaran les funcions de validació, indexació, catalogació, establiment de freqüència de captura, control de duplicats i allotjament. Els programes que realitzaran aquestes tasques seran Nutch Wax, Heritrix i Web Curator.

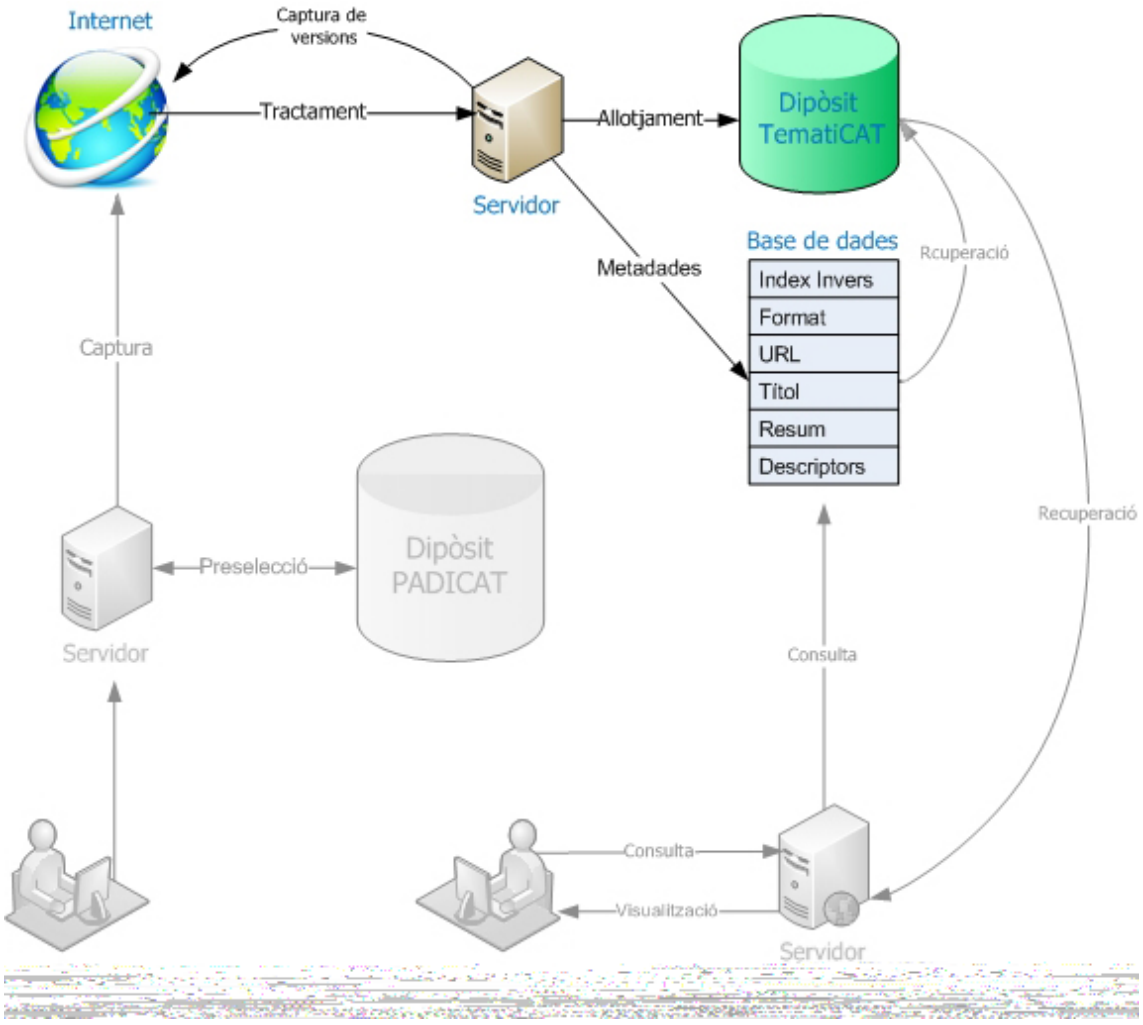


Fig. 27. Detall del sistema de tractament dins el sistema global

En l'apartat de la implementació del mòdul d'entrada ja s'han expressat els motius i els avantatges que suposa seleccionar els programes Heritrix i Nutch Wax. Pel que fa a l'aplicació Web Curator, aquesta s'ha seleccionat per la possibilitat de realitzar múltiples tasques amb una única interfície d'administració. A banda de la seva polivalència, que permet satisfer les necessitats expressades en aquest projecte, Web Curator és plenament compatible amb la resta de programes que formen el sistema, i compte amb el recolzament del IIPC. A partir de la instal·lació d'aquests programes, l'arquitectura del mòdul de tractament i les seves funcionalitats queden definides de la següent manera:

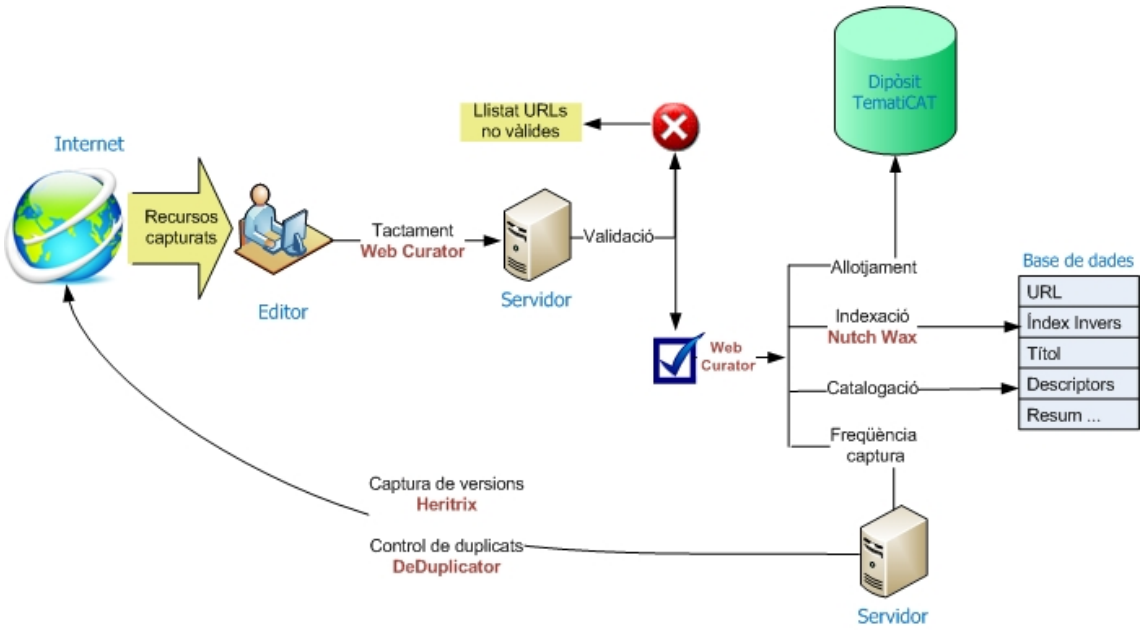


Fig. 28. Sistema de tractament

5.3.1. Implementació i adaptació de Nutch Wax (Indexació)

El programa Nutch Wax ja haurà estat implementat en la fase del mòdul de d'entrada de recursos. En aquest estadi de la instal·lació l'objectiu serà la configuració de la funció d'indexació de recursos, que tindrà com a finalitat la creació d'un fitxer invers dels termes significants del document.

Inicialment s'haurà de definir quines paraules no hauran de ser indexades per la seva manca de significat (*stopwords*), i llavors, s'haurà de crear un camp en la base de dades que estarà destinat a recollir els termes resultants de la indexació. En una vessant més tècnica, s'haurà d'estudiar la necessitat de realitzar reindexacions per mantenir la consistència d'aquests índexs.

Es realitzaran proves pilot a partir de la indexació de col·leccions amb diferent nombre de documents, per tal d'observar quina és la durada d'aquesta tasca.

5.3.2. Implementació i adaptació de Web Curator

La implementació de l'aplicació Web Curator té prevista una duració aproximada de 60 hores. A banda de la instal·lació del programa en el servidor, s'haurà de crear una base de dades MySQL que contindrà les metadades dels recursos allotjats en el dipòsit.

Una vegada finalitzada la implementació del programa s'haurà de supervisar que les funcionalitats que realitzarà aquest mòdul responen correctament a les necessitats establertes. Per garantir el correcte funcionament d'aquesta part del sistema es faran una sèrie de proves pilot amb alguns dels recursos capturats en la fase anterior:

- Validació de recursos: just abans d'entrar en el circuit de tractament de recursos, s'haurà de comprovar que el sistema permet rebutjar els elements que no compleixen els criteris predefinits de la col·lecció. Del conjunt de recursos que no superin la validació s'haurà de generar un llistat d'URLs per tal d'evitar que siguin una altra vegada capturats.

- Catalogació de recursos: per tal de facilitar l'agregació de descriptors als recursos, es crearà un camp que actuarà com a llista de validació, el qual contindrà la totalitat de matèries que es podran fer servir per catalogar una col·lecció concreta[91]. En el moment de descriure un determinat web, només s'hauran de seleccionar les matèries pertinents, i aquestes quedaran gravades en el registre de la base de dades. En la prova pilot d'aquesta funció es descriuran un centenar de registres amb la finalitat de comprovar que les dades queden correctament introduïdes en els corresponents registres de la base de dades, els quals seran aprofitats per poder sotmetre a prova la interfície de consulta en la següent fase d'implementació.

- Establiment de freqüència de captura: a patir d'un conjunt de webs es definiran diferents períodes de captura a través del Web Curator. En aquesta funcionalitat s'hauran de configurar els paràmetres de quan ha de començar la captura de cada recurs i cada quan s'ha d'executar aquesta acció. Una vegada s'hagin definit els criteris de la captura de versions posteriors dels recursos, s'haurà de comprovar la comunicació entre el programa Web Curator (on s'hauran definit les ordres) i Heritrix (encarregat d'executar la captura). Juntament amb les proves pilot que es realitzin de la captura periòdica programada de webs, s'analitzarà el correcte funcionament del mòdul de control de duplicats d'Heritrix, DeDuplicator, que ja haurà estat implementat dins el marc de la fase d'ingestió de recursos. Per poder portar a terme aquesta prova pilot s'hauran de definir una sèrie de paràmetres en el mòdul DeDuplicator per tal de definir els criteris de comparació dels arxius web.

[91] Degut a que la creació d'aquesta llista en el *framework* del programa Web Curator requereix un cert desenvolupament tecnològic, en el cas de presentar-se problemes tècnics que puguin incrementar la duració de la implementació, es preveu la utilització d'una base de dades desvinculada del *software*.

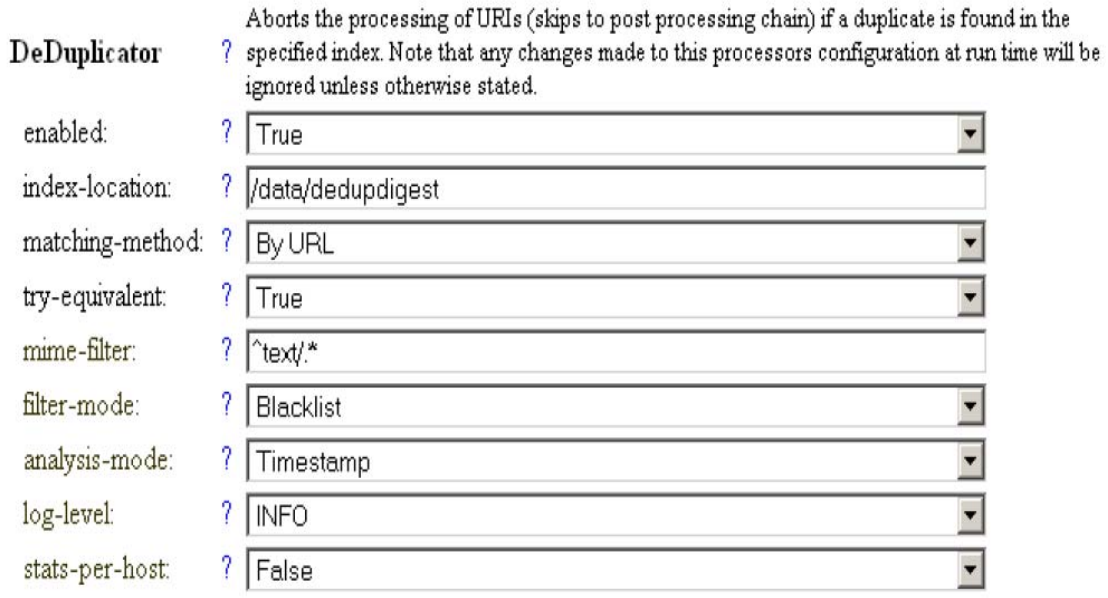


Fig. 29. Paràmetres de configuració del mòdul de control de duplicats d'Heritrix

- Allotjament: una vegada finalitzada tota la fase de tractament dels recursos digitals, s'haurà de comprovar que els arxius queden correctament allotjats en el dipòsit per a la seva posterior recuperació.

5.4. Implementació del mòdul de recuperació i visualització de recursos

El mòdul de recuperació i visualització és la part del sistema mitjançant la qual els usuaris podran accedir als recursos allotjats en el dipòsit. Els programes que faran possible aquest accés són Nutch Wax (consulta) i Wera (visualització).

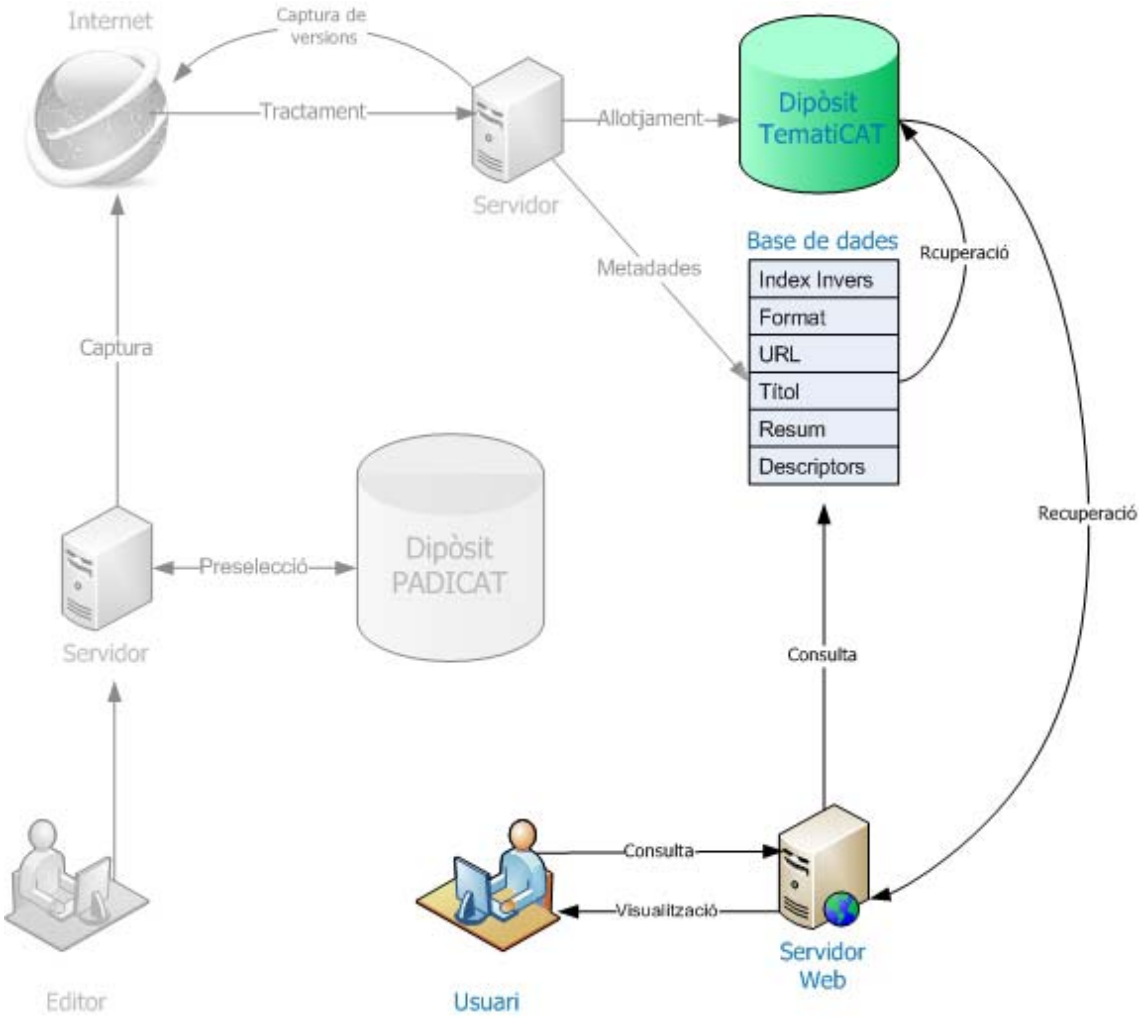


Fig. 30. Detall del sistema de recuperació i visualització dins el sistema global

Pel que fa a l'elecció del programa Wera, s'han tingut en compte dos tipus de factors. En primer lloc s'han valorat els aspectes de compatibilitat, ja sigui amb el PADICAT perquè també usa aquesta aplicació, o bé amb la resta de programes del sistema, ja que hi ha altres projectes que destaquen la plena integració entre Wera i Nutch Wax. En segon lloc s'han valorat més positivament les prestacions de Wera que les de Wayback machine, ja que el sistema de cerca permet interrogar obertament amb paraules clau i amb altres camps de descripció, i també, a nivell de visualització, per la forma d'oferir els resultats i les possibilitats de navegació que disposa. L'arquitectura i funcions d'aquest mòdul queden distribuïdes de la següent manera:

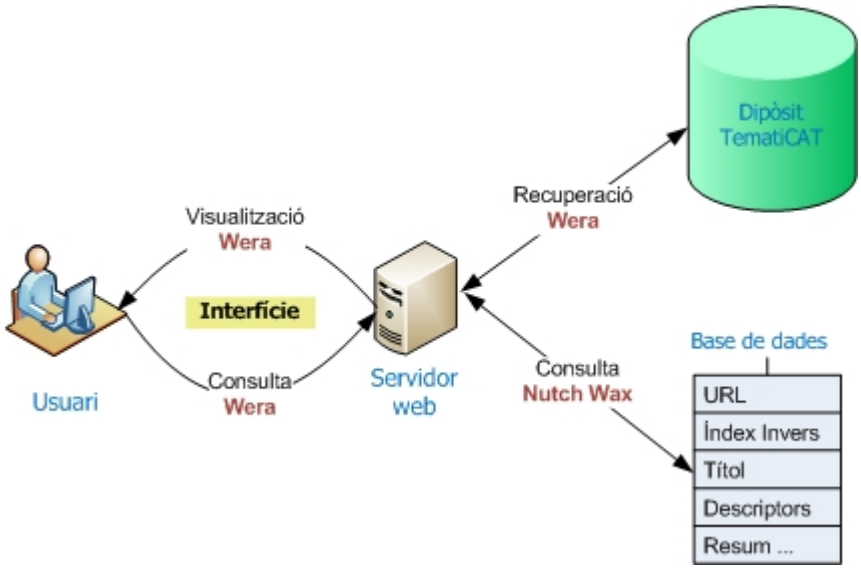


Fig. 31. Sistema de recuperació i visualització

5.4.1. Implementació i adaptació de Wera

La implementació de l'aplicació Wera té prevista una duració aproximada de 60 hores. La configuració de les tasques de recuperació i visualització de Wera es realitzaran mitjançant una interfície provisional de consulta, ja que en aquest projecte no es contempla ni el disseny ni la implementació de la interfície final de l'usuari. En els següents apartats es mostren els detalls de la instal·lació de Wera i les proves pilots previstes per comprovar el seu correcte funcionament:

- Consulta de recursos: es configurarà l'aplicació Wera perquè pugui enviar consultes al motor de cerca de Nutch Wax, concretament als camps de descripció URL, matèria, data i col·lecció, així com al fitxer invers (per interrogacions a text lliure). Es faran proves combinant les diferents eines de cerca i també es comprovarà la possibilitat de fer servir els operadors booleans i de truncament.
- Visualització: s'elaborarà un patró de visualització dels registres resultants de la cerca efectuada en el dipòsit, on es mostrarà la següent informació de cada recurs: títol, URL i un fragment de text del web. Es configurarà l'aplicació perquè permeti visualitzar el web en qüestió, veure més recursos pertanyents als mateix domini i expandir el registre, que significa que es podrà tenir accés a totes les dades catalogràfiques (metadades i llista de versions del lloc web).
- Recuperació de recursos: es configurarà el mòdul de recuperació de Wera (*Document retriever*) perquè visualitzi els arxius allotjats en el dipòsit que l'usuari indiqui. Ens aquesta tasca es faran proves de recuperació de diferent tipologia de

webs (estàtiques, dinàmiques, etc.), per observar si es mostra algun tipus de problemàtica a la fi de ser corregits.

6. Col·lecció temàtica Beta: Eleccions municipals a Catalunya 2007

Amb la finalitat de realitzar una prova pilot del funcionament global del sistema d'informació del TematiCAT es presenta la planificació d'una col·lecció temàtica basada en la celebració de les eleccions municipals a Catalunya el 27 de maig del 2007. L'elaboració d'aquesta col·lecció ha de permetre analitzar el nivell d'assoliment dels principals processos del projecte:

- Adaptació del sistema de selecció i captura
  - ► Eficàcia de les estratègies de selecció de recursos
  - ► Duració de les captures dels recursos seleccionats
- Adaptació del sistema de tractament dels recursos
  - ► Duració de la indexació dels recursos validats
  - ► Validesa de les estratègies de descripció dels recursos
  - ► Funcionament del mòdul automàtic de captura periòdica
  - ► Eficàcia del mòdul de control de duplicats
- Adaptació del sistema d'accés als recursos
  - ► Eficiència en la recuperació dels recursos
  - ► Visualització dels recursos
  - ► Capacitat de navegació de les versions capturades dels recursos

L'anàlisi de tots aquests aspectes, entre d'altres més específics, a través dels resultats que s'obtindran amb la creació de la col·lecció temàtica beta, oferirà una radiografia de la consistència del sistema que permetrà poder realitzar les modificacions necessàries abans de donar per acabat el projecte.

Abast de la col·lecció

Segons les definicions donades en l'apartat de selecció i captura de recursos, la col·lecció en qüestió que s'està planificant pertanyeria clarament al conjunt d'esdeveniments programats, ja que unes eleccions tenen una periodització molt estricte, la qual es fa publica amb àmplia anticipació a les dates de la seva celebració[92].

L'ambició d'aquesta col·lecció és que estigui formada per una quantitat de llocs webs que estigui al voltant dels 1.500[93], més les seves respectives versions, el nombre del qual variarà segons el grau d'actualització de cadascun.

Abast temàtic

Els recursos seleccionats per formar part de la col·lecció de les eleccions municipals catalanes de 2007, han d'estar estrictament relacionats amb l'àmbit polític català i la celebració d'aquests comicis. Tenint en compte aquesta definició general de l'abast temàtic de la col·lecció, s'entendrà que qualsevol afer social, econòmic o cultural que s'hagi explícitament relacionat amb la cursa electoral serà susceptible de ser inclòs.

Abast temporal

El límits cronològics de la col·lecció, que marcaran la fase de captura de recursos web, vindran donats pel calendari electoral esmentat a l'inici d'aquest apartat, en el qual es defineixen totes les activitats vinculades amb la celebració dels comicis municipals. L'abast temporal de la col·lecció estarà comprès entre l'1 d'abril, moment en que es proclamen les candidatures, i el 30 de juny, data en la que els ajuntaments ja han d'haver estat constituïts.

Abast geogràfic

L'abast territorial de la col·lecció es centrarà a Catalunya en general, i que específicament intentarà donar cobertura a tots els municipis catalans

[92] GENERALITAT DE CATALUNYA. DEPARTAMENT DE GOVERNACIÓ I ADMINISTRACIONS PÚBLIQUES. *Calendari electoral de les eleccions municipals 2007* [En línia]. Barcelona: Generalitat de Catalunya, febrer de 2007. [Data de consulta: 23/03/2007]. Disponible a: <http://www.gencat.net/governacioap/eleccions/LOCALS 2007/MUNICIPALS07/m07_cale.pdf>.

[93] Respecte al nombre de recursos que formaran la col·lecció s'ha de tenir en compte que uns 900 seran els llocs web dels ajuntaments de Catalunya.

Tipologia dels recursos

La tipologia dels recursos que integraran la col·lecció beta s'ha d'enfocar des de la vessant tècnica i la vessant de continguts.

Pel que fa als aspectes més tècnics, s'han de definir les característiques dels llocs web que podran ser acceptades. Tal com s'ha expressat en els corresponents apartats d'anàlisi i disseny, el sistema no ha de tenir cap problema per gestionar webs estàtiques i dinàmiques. En tot cas, seguint les recomanacions d'altres projectes, s'evitarà capturar webs protegides per *passwords*, que tinguin bases de dades en obert o que tinguin programes incrustats del tipus calendari. Pel que fa als formats, es prioritzarà la captura i conservació dels recursos que estiguin construïts amb estàndards.

A nivell de continguts, sempre que siguin seriosos, no es farà cap tipus de restricció per ser inclòs en la col·lecció. Malgrat aquesta premissa, es prioritzarà la captura de recursos relacionats amb fonts oficials de l'administració pública, partits polítics i entitats afins, i mitjans de comunicació.

Timing

Està previst que la duració de la prova pilot de la col·lecció temàtica de Eleccions municipals a Catalunya 2007 sigui de quatre mesos, entre la segona quinzena de març i la primera quinzena de juliol. En la taula següent es desglossa la periodificació de cada fase d'aquesta prova pilot:

Període / ***Tasques***	**Març**	**Abril**	**Maig**	**Juny**	**Juliol**
Planificació de la col·lecció					
Preselecció dels recursos					
Captura dels recursos					
Tractament dels recursos					
Proves d'accés als recursos					
Modificacions					

Fig. 32. Timing de la fase de creació de la col·lecció temàtica Beta

6.1. Selecció i captura dels recursos

Per les característiques de la col·lecció temàtica que es pretén crear, es seguirà l'estratègia de selecció i captura de recursos definida per la tipologia d'esdeveniments programats. Resumint el que s'ha explicat en el seu corresponent apartat, aquesta estratègia es basa en una preselecció de recursos obtinguda del fons del dipòsit PADICAT, que ha de servir per localitzar els recursos especialitzats o relacionats amb la temàtica de la col·lecció. En funció de quan s'hagi portat a terme l'última captura integral del PADICAT s'obtindran més o menys recursos que continguin informació relacionada amb les eleccions municipals de 2007, però en tot cas, tots els llocs webs vinculats amb l'àmbit polític, els quals formen part de l'abast de la col·lecció, seran identificats per mitjà d'aquesta preselecció. S'estima que més d'un 75% dels llocs web que formaran la col·lecció seran identificats en la fase de preselecció.

Tenint en compte els recursos que interessa que formin part d'aquesta col·lecció, abans d'executar la cerca de webs potencialment interessants en el dipòsit PADICAT, es crearà una llista de llocs que es podran recuperar fàcilment de forma manual, amb la finalitat de reduir al màxim els costos tecnològics. Aquesta llista estarà formada per les adreces de tots els ajuntaments, diputacions i partits polítics que disposen de lloc web, les quals s'han trobat agrupades en un document[94].

Per tal de formar aquesta preselecció de webs disposem de dues estratègies de cerca que s'executaran mitjançant el motor de cerca de Nutch Wax: consulta al camp de descripció de matèria i consulta a text lliure. En el cas de la primera, s'analitzaran quins descriptors s'han fet servir per catalogar els recursos que cobreixen la temàtica política en el PADICAT, i partir d'un llistat d'aquests es realitzaran diverses interrogacions, fins observar que els resultats són redundants a altres obtinguts[95]. Pel que fa a la cerca a text lliure, es proposaran una sèrie de termes i noms propis relacionats amb la temàtica de la col·lecció, i a partir d'aquests, mitjançant la lògica booleana es crearan equacions de cerca. Una llista dels possibles termes que es faran servir en aquesta estratègia de cerca són els següents:

[94] MUNICIPIS I COMARQUES DE CATALUNYA. *Sistema d'informació d'administració local* [En línia]. Catalunya: MUNICAT, 2007. [Data de consulta: 22/03/2007]. Disponible a: <http://www.municat.net/pagines/descarregues/desc.htm>.

[95] Respecte a la cerca en el camp de matèria, s'haurà de tenir en compte que el PADICAT encara no està en una fase 100% operativa, i que per tant, actualment no té tots els recursos del dipòsit catalogats. Tot i així, segons els plantejaments del projecte, es preveu que es realitzi la descripció de matèria de tots els recursos, cosa que fa viable a curt termini l'estratègia de cerca a través d'aquest camp.

Descriptor	Observació
Eleccions	Descriptors temàtics
Comicis	
Electoral	
Campanya	
Ajuntament	
Política	
Polític	
Alcalde / Batlle	
Diputació	
“27 de maig”	Descriptors cronològics
2007	
Catalunya	Descriptors geogràfics
Barcelona	
Girona	
Lleida	
Tarragona	
PSC / Partit Socialista de Catalunya	Nom dels partits polítics amb més representació a Catalunya
CIU / Convergència i Unió	
ERC / Esquerra Republicana de Catalunya	
ICV / Iniciativa per Catalunya els Verds	
PPC / Partit Popular de Catalunya	
C’s / Ciutadans	
PXC / Plataforma per Catalunya	
Jordi Hereu	Noms propis dels principals candidats per Barcelona
Jordi Portabella	
Imma Mayol	
Xavier Trias	
Alberto Fernández Díaz	
Anna Pagans	Noms propis dels principals candidats per Girona
Josep Maria Jofre	
Cristina Alsina	
Joan Olóriz	
Concepció Veray	
Àngel Ros	Noms propis dels principals candidats per Lleida
Isidre Gavín	
Ismael Zapater	
Xavier Sáez	
Mercè Rivadulla	
Josep Ballesteros	Noms propis dels principals candidats per Tarragona
Joan Aregio	
Alejandro Fernández	
Sergi de los Rios	
Josep Montilla	Noms d’altres personalitats polítiques relacionades amb la precampanya
Artur Mas	
Josep Lluís Carod Rovia	
Josep Piqué	
Joan Saura	
José Luís Rodríguez Zapatero	
Mariano Rajoy	

Fig. 33. Paraules clau usades en les estratègies de cerca

En l'execució de qualsevol de les dues estratègies de cerca proposades, es programarà el motor de cerca, mitjançant les URLs, perquè obviï els llocs web detectats manualment, a la fi de no haver de gestionar dues vegades els mateixos recursos.

Els recursos obtinguts en la cerca del dipòsit del PADICAT seran validats per comprovar la pertinença en l'abast de la col·lecció, i llavors, juntament amb el conjunt de webs seleccionades manualment, se'n capturarà la versió vigent penjada a Internet mitjançant el programa Heritrix. Aquest grup de llocs web, que es preveu que superi àmpliament el miler, passarà directament a la fase de tractament on seran indexats, catalogats i se'n determinarà la freqüència de captura.

En una col·lecció amb una temàtica tant popular com la que es presenta, cap la possibilitat que la preselecció sigui tant exhaustiva que s'arribi al còmput de recursos estimats només amb aquesta fase. En el cas que no sigui així, es passarà a la següent fase de l'estratègia de selecció i captura d'esdeveniments programats: la selecció directe a Internet.

En aquesta segona fase de selecció i captura es podran fer servir dues estratègies de cerca: la cerca amb filtres i la cerca a través d'algoritmes. En el cas de la segona, es descartarà directament, ja que el gruix de recursos que s'hauran obtingut amb les altres estratègies i la inversió que requereix crear un algoritme de cerca no compensa el nombre de recursos nous que pot oferir.

Pel que fa la cerca amb filtres, aquesta es realitzarà a través de l'aplicació Heritrix, que haurà de ser programada amb els següents paràmetres:

1. Primer de tot perquè pugui començar a navegar se li hauran de donar unes coordenades, que seran les URLs dels dominis capturats pel PADICAT els dos últims anys (sense tenir en compte les versions).
2. De les URLs seleccionades del PADICAT, s'hauran d'extreure les que ja han estat preseleccionades per tal de no duplicar esforços.
3. Es limitarà el temps de descàrrega i el nombre d'arxius en funció de les dades obtingudes en les proves pilot de la implementació
4. Es configurarà el robot perquè visiti els enllaços i segueixi navegant a través d'ells.
5. Es configurà la profunditat de captura dels recursos filtrats a la totalitat del lloc web mentre no superi les limitacions establertes.

6. Es configurarà Heritrix perquè capturi només els webs que tinguin en la zona del títol o en els encapçalaments una sèrie de termes[96] que li seran donats. Aquesta cerca filtrada es pot realitzar en diferents treballs paral·lels per tal de combinar termes.

7. Es configurarà Heritrix perquè descarregui els recursos en format comprimit i els allotgi en l'espai de treball del dipòsit TematiCAT, per tal de passar a la fase de tractament.

6.2. Tractament dels recursos

La fase de tractament començarà una vegada s'hagi realitzat la preselecció de recursos, i a partir d'aquest estadi s'executarà simultàniament al procés de captura. Aquesta és la fase que presenta menys automatització ens els processos, ja que requereix la intervenció de l'agent humà en el procés de validació, catalogació i en l'establiment de freqüència de captura. Per tal de fer una estimació dels recursos necessaris per poder tractar tots els webs de la col·lecció, s'han realitzat unes simulacions que determinen que la duració de les tasques de validació, catalogació i establiment de freqüència està al voltant del 15 minuts per lloc web. Si partim de la base que el projecte pretén gestionar 1.500 llocs web (sense comptar les versions), el tractament d'aquest fons suposa 375 hores aproximadament de treball.

El cicle de tractament de les versions dels recursos originals es computa apart, ja que no requereix el mateix temps. Inicialment el control duplicats és una tasca automàtica, i la validació només requerirà un anàlisi superficial, ja que es parteix de la suposició que les captures periòdiques són a priori potencialment interessants. Pel que fa a la catalogació, només s'haurà de revisar que els descriptors heretats del recurs original encaixin en la nova versió, i en cas que sigui necessari es podran substituir. No es pot fer una valoració exacte de la duració, ja que és impossible saber quantes versions es capturaran. Per tal de fer una estimació dels recursos necessaris es posa com a xifra assolible els 10.000 lloc webs, comptant les originals i les versions. Novament mitjançant una simulació, considerarem que per processar totalment en la fase de tractament una versió d'un web es necessitaran uns 5 minuts, és a dir, que per tractar la totalitat de les captures periòdiques es necessitaran 710 hores de treball d'un agent especialitzat en gestió de recursos digitals.

[96] Vegeu la taula de la llista de descriptors de la pàgina 107

Validació

El procés de validació és l'últim filtre que ha de passar un recurs per tal de considerar-se apte per formar de la col·lecció. Per tal de seguir una coherència a l'hora de decidir la seva validesa es seguirà el següent patró:

- Temes concrets relacionats amb les eleccions
    - Política urbanística
    - Política social
    - Política mediambiental
    - Política cultural
    - Política sanitària
    - Polèmiques polítiques
    - Infraestructures
    - Seguretat ciutadana
    - Immigració
- Àrees geogràfiques (Catalunya)
    - A nivell municipal
    - A nivell comarcal
    - A nivell provincial
    - A nivell de l'autonomia
- Períodes cronològics
    - 1 d'abril – 30 de juny
- Personalitats
    - Polítics
    - Empresaris
    - Intel·lectuals
    - Celebritats
- Entitats relacionades amb l'esdeveniment:
    - Institucions públiques
    - Institucions culturals
    - Institucions d'estudis social i polítics
    - Empreses
- Longitud dels continguts
    - Mínim dos nivells verticals i dos horitzontals en relació a l'estructura del web
    - No hi haurà limitacions de màxims

- Tipologia de documents prioritaris:
    - Text: html, text i pdf
    - So: mp3
    - Imatge: jpeg i gif
    - Imatge en moviment: mpeg i avi

Els recursos web que superin la validació passaran als següents processos de tractament. Per altra banda, els que no la superin alimentaran la llista de recursos no pertinents a la col·lecció, per tal que no siguin capturats una altra vegada.

Indexació

La indexació dels recursos validats es farà de forma automatitzada mitjançant l'aplicació Nutch Wax, la qual crearà un fitxer invers amb les paraules significants de cada recurs per tal de proporcionar un punt d'accés a la recuperació. La indexació serà una tasca que s'executarà immediatament una vegada el recurs hagi estat validat ja que és un dels processos que té una duració més llarga.

Catalogació

La catalogació dels recursos té dos vessants tècniques: l'automatitzada i la manual.

En el cas de la catalogació automatitzada es realitza a través del mòdul d'indexació de Nutch Wax, el qual extraurà una sèrie de metadades que seran integrades als registres de la base de dades dels recursos. A través d'aquest procés automatitzat, el qual ha estat prèviament parametritzat, obtindrem les dades de les següents metadades:

- Identificador
- Títol
- Llengua
- Format de l'arxiu
- Data de captura
- Resum

La catalogació manual serà realitzada per un agent editor especialitzat en la descripció de documents. Tal com s'ha explicat en els anteriors apartats, per les característiques que tenen els recursos digitals, només s'emprarà el nivell mínim de descripció del Dublin Core, però prescindint del camp autor o contribuïdor per tal d'evitar problemes de consistència en la catalogació. Per tant, el documentalista encarregat d'aquesta

tasca únicament haurà de descriure el camp matèria, a través de l'aplicació Web Curator, i usant com a referència la LEMAC. Tot seguit es presenta un exemple dels descriptors que es podran fer servir en la catalogació:

- Alcaldes
- Candidatures
- Documents electorals
- Dret electoral
- Eleccions locals
- Finances municipals
- Funcionaris i empleats municipals
- Govern i administració
- Governs de coalició
- Municipis
- Participació política
- Partits polítics
- Política
- Política d'habitatge
- Política urbana
- Premsa política
- Propaganda electoral
- Regidors municipals
- Seguretat ciutadana
- Serveis municipals
- Sòl, Ús urbà del
- Transport urbà
- Urbanisme
- Vot

Freqüència

L'establiment de la freqüència de captura dels recursos validats també es realitzarà a través de l'aplicació Web Curator, que executarà el programa Heritrix per recopilar les versions dels llocs web. Com en els anteriors processos explicats, es crearan unes pautes per decidir amb quina freqüència es capturarà cada web. Aquestes pautes es definiran pel tipus de web, en funció de la qual s'establirà una periodicitat de captura més o menys àmplia. Tot seguit s'exposa una proposta dels períodes de descàrrega en base a una classificació dels webs que conformaran la col·lecció:

▪ Mitjans de comunicació	1 captura diària
▪ Ajuntament de municipis amb més de 100.000 habitants	2 captures setmanals
▪ Ajuntament de municipis amb més de 20.000 habitants	1 captura setmanal
▪ Ajuntament de municipis amb menys de 20.000 habitants	2 captures mensuals
▪ Partits polítics amb representació parlamentària	2 captures setmanals
▪ Partits polítics sense representació parlamentària	2 captures mensuals
▪ Webs personals de personatges de relleu polític	1 captura setmanal
▪ Institucions públiques (diferents als ajuntaments)	1 captura setmanal
▪ Fundacions polítiques i de recerca	1 captura setmanal
▪ Blocs o webs d'opinió	2 captures setmanals
▪ Altres webs	A decidir pel documentalista

Les freqüències de captura estaran subjectes a modificacions en funció del nivell d'actualització que presentin els webs. Pel que fa a la captura, al tractar-se d'una

operació automatitzada, es programarà Heritrix perquè les realitzi a la matinada, ja que la xarxa en aquestes hores té menys tràfic, facilitant les tasques de descàrrega d'arxius. Igual que la captura, el control de duplicats serà una tasca que es realitzarà de forma automàtica a través del mòdul DeDuplicator d'Heritrix.

L'última tasca que es realitzarà en la fase de tractament serà l'allotjament definitiu dels recursos que formaran la col·lecció en el dipòsit del TematiCAT.

6.3. Accés als recursos

Amb el sistema de tractament processant els recursos capturats s'anirà conformant la col·lecció temàtica de les eleccions municipals a Catalunya de 2007. Mitjançant aquesta última fase de la prova pilot, es comprovarà que es poden recuperar i visualitzar els ítems allotjats en el dipòsit usant les diferents eines de cerca disponibles i navegant a través dels diversos punts d'accés que s'habilitin.

En relació a les eines de cerca es faran consultes sobre conjunts d'informació, dels quals en sabrem el resultat a priori per tal de comparar-lo amb el llistat obtingut de la interrogació al dipòsit. En el test de la consulta s'hauran de comprovar els següents camps per tal de confirmar el seu correcte funcionament.

- **Paraula clau**: s'hauran de realitzar cerques amb més d'un terme per comprovar el funcionament dels operadors booleans i de truncament.
- **Adreça web**: es faran proves amb diverses URLs a finalitat d'observar que es poden recuperar correctament els recursos de forma individual
- **Matèria**: es faran cerques amb tots els descriptors disponibles.
- **Dates**: per comprovar aquest camp es repetiran les anteriors cerques afegint una limitació cronològica.
- **Ordenar**: es comprovarà que es poden realitzar ordenacions ascendents i descendents dels resultats obtinguts en les cerques anteriors.

Pel que fa a la visualització s'haurà de comprovar que els recursos són recuperats correctament i que la navegació dins dels llocs webs sigui possible. A nivell de les prestacions que ofereix Wera, s'hauran de realitzar proves sobre l'opció de veure totes les versions d'un recurs concret i la possibilitat de veure les dades catalogràfiques d'un web a través de l'opció d'expandir registre.

L'última tasca a realitzar en aquesta prova pilot serà la creació i comprovació dels punts d'accés alternatius que ocuparan els quatre mòduls de la interfície. En relació a aquesta col·lecció temàtica es proposen els següents reculls específics de recursos:

- Els partits polítics: des d'aquest mòdul es podrà accedir a un llistat dels webs dels partits polítics catalans que han estat objecte de captura.
- Els ajuntaments de Catalunya: en aquest mòdul es podrà trobar un llistat alfabètic de tots els ajuntaments de Catalunya que disposen de pàgina web, el qual estarà vinculat amb les seves corresponents captures allotjades en el dipòsit.
- Les notícies més destacades: en funció de les notícies que vagin sorgint al llarg de la campanya electoral, aquestes es poden reflectir individualment o en conjunt en aquest mòdul. Per tal de relacionar-ho amb els recursos del dipòsit, es mostraran una llista de webs, principalment de mitjans de comunicació, que mostren diferents punts de vista sobre els fets que es volen destacar.
- Webs proposades per usuaris: En aquest mòdul s'hi ubicarà la llista de webs que hagin estat suggerides pels usuaris del TematiCAT. Tot i que aquest mòdul no serà operatiu en aquesta prova pilot, ja que no es podrà comptar amb la participació dels usuaris al no estar el sistema en línia en obert, s'inclou en aquesta llista per la importància de plasmar la col·laboració de la comunitat internauta.
- Webs més visitades: com en l'anterior cas, aquest mòdul només seria operatiu al cap d'uns dies que la col·lecció hagi estat posada en obert, i es pugui fer un anàlisi dels *logs* i les paraules clau més utilitzades en les cerques. En aquest mòdul pot funcionar com del *top 10* del dipòsit, posant setmanalment les 10 webs més visitades per part dels usuaris del projecte.

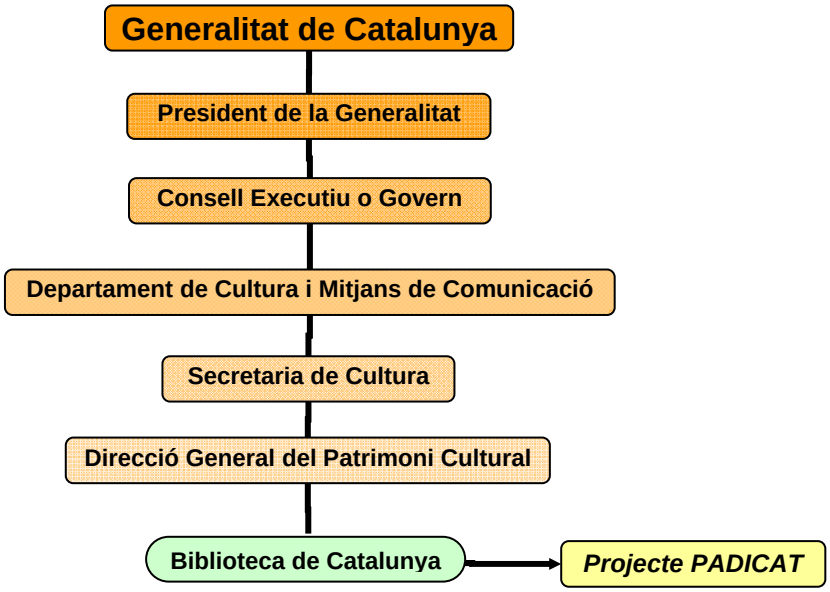
Modificacions

A partir de totes la prova pilot realitzada per la creació de la col·lecció temàtica de les Eleccions municipals a Catalunya 2007 s'avaluarà el rendiment del sistema i els resultats obtinguts. D'aquesta avaluació se n'extraurà un informe de les modificacions necessàries per aconseguir un òptim funcionament del sistema, les quals hauran de ser realitzades, juntament amb proves pilot puntuals, per donar per acabat el projecte.

Annexos

Sumari d'annexos

Annex 1. Organigrama de dependència de la Biblioteca de Catalunya



Annex 2. Característiques dels models de dipòsits Nacionals de recursos digitals

Model integral

Punts forts del model integral

- **Exhaustivitat i riquesa de la col·lecció, en quantitat i qualitat:** aquest model permet fer una captura gairebé completa d'un instant del web, el que s'anomena un *snapshot*[97] (sempre tenint en compte els criteris d'inclusió). D'aquesta manera s'evita entrar en el debat de la selecció de recursos, i per tant no haver d'entrar en el terreny de quins són els recursos que són convenients i quins no. Internet ha evolucionat molt des dels primers dies i és gairebé impossible saber quins poden ser els interessos públics i els àmbits que s'investigaran en els propers anys. Alhora, el fet de fer una captura completa permet respectar més fàcilment la interrelació que hi ha entre els recursos, i per tant poder mantenir dins el dipòsit l'estructura transversal que hi ha entre els recursos.

- **Compilació automàtica:** el model integral funciona per mitjà d'uns robots, *crawlers*, que executen de forma automàtica la fase de captura. L'esforç en aquest sentit rau en la dificultat de programar aquests buscadors, ja que s'han d'elaborar uns paràmetres molt específics per evitar el soroll o el silenci d'informació o resultats. Una vegada construït l'algoritme de cerca que mou el robot es pot executar automàticament en els períodes establerts, permeten en cas que fos necessari modificacions d'aquesta equació per poder refinar la cerca. Això significa que els costos de la captura es minimitzen substancialment en relació el model selectiu, i que alhora es pot executar una captura total de la web amb un lapse de temps molt curt.

- **Baix cost:** observant els aspectes anteriorment esmentats es fa palès que aquest model, i en relació amb el selectiu, redueix els costos sobretot en l'apartat de recursos humans. Mentre que en els projectes on es fa selecció de recursos hi treballen una mitja superior a 10 persones, mentre que en els que estan automatitzats la xifra és clarament inferior.

[97] El concepte *snapshot* s'empra per definir la captura que es fa de la web en un període concret, que sol ser de dos a quatre setmanes en funció dels criteris establerts. Per tant, els *snapshots* són les *fotografies* periòdiques que s'obtenen de la web al executar la fase de captura.

Punts febles del model integral

- **Impossibilitat d'accedir a la Internet Invisible**: un dels principals problemes dels sistemes automatitzats és que només tenen accés als recursos publicats en règim obert, és a dir que tenen vetat l'accés a la Internet Invisible, la qual s'ha calculat que multiplica pel cap baix entre dos i cinquanta vegades la Internet visible[98]. Per tant es deixen de capturar tots aquells recursos que són de pagament, que estan codificats sota *password* i la major part de pàgines dinàmiques. Més que un problema de quantitat, és un greu problema de qualitat, ja que aquestes solen ser les pàgines més riques en continguts i informació de rigor. Avui en dia s'està intentant superar aquest obstacle per mitjà d'acords directes amb les organitzacions productores d'aquests recursos que formen part de la Internet Invisible, sent aquest un dels motius que faci tendir cap al model híbrid.

- **Compilació irregular de la col·lecció**: el model integral té establert un calendari periòdic de captures que no té en compte la publicació o actualització dels recursos. Per aquest motiu es generen llacunes en les col·leccions, les quals només es poden evitar establint períodes concrets per determinats recursos, implicant un increment de la complexitat del sistema.

- **Accés limitat als resultats**: la gran quantitat de recursos que es capturen en el marc del projectes que segueixen el model integral fa gairebé impossible que es puguin catalogar, això dificulta la seva recuperació. Per altra banda, la manca d'un text legal que defineixi l'àmbit de difusió d'aquests recursos per part dels organismes responsables de la captura, fa que aquest només siguin consultables en les seves instal·lacions. La majoria de projectes liderats per biblioteques nacionals segueixen aquesta fórmula de difusió. Internet Archive és la gran excepció fent accessible tota la seva col·lecció de recursos digitals en línia.

[98] Aguillo, Isidro. *Internet Invisible* [En línia]*: los contenidos son la clave*. InternetLab : CINDOC-CSIC, abril 2003. [Data de consulta: 08/05/2006]. Disponible a: < http://internetlab.cindoc.csic.es/cursos/Internet_ Invisible2003.pdf>.

Model selectiu

Punts forts del model selectiu

- **Creació d'una col·lecció equilibrada**: aquest és un aspecte relacionat amb el més estricte àmbit bibliotecari tradicional, ja que la col·lecció es construeix a través del procés d'adquisició mitjançant un agent que en fa la selecció, descripció i una posterior actualització en el cas que sigui necessari. La composició de la col·lecció es sol vertebrar al voltant d'un eix temàtic i que habitualment es desglossa en forma de taxonomia. L'equilibri de la col·lecció ve donat pel nombre i el coneixement dels recursos que es posseeixen a través de la descripció i la seva categorització.

- **Facilitat d'accés al fons:** la catalogació dels recursos permet una recuperació eficient, possibilitant consultes complexes a través de diferents camps de descripció. Aquesta característica dóna una gran versatilitat als recursos, oferint així la possibilitat d'integració d'aquest en un sistema de catàleg bibliotecari.

- **Estratègic:** el fet de funcionar amb un sistema d'acords amb els organismes productors permet en la majoria de casos poder pactar la difusió en línia dels recursos, i alhora aconseguir les dades dels editors i dels recursos per ser catalogats més fàcilment. També mitjançant aquest sistema d'acords és possible capturar una part de la Internet Invisible, ja que es garanteixen beneficis per ambdues parts.

Punts febles del model selectiu

- **Parcialitat en descriure el món:** la selecció de recursos implica un judici subjectiu del seu valor i una anticipació del que serà d'interès públic en el futur. Els recursos que es seleccionen solen ser una mostra molt petita davant la captura que es realitza en el model integral. Això significa que es perden molts documents que poden significar altres punts de vista, ampliacions d'un tema, matisos o simplement un aspecte que no ha semblat interessant en un moment determinat, per tant la realitat des del món digital no es copsa fidelment.

- **Cost elevat:** com ja s'ha anticipat en l'apartat anterior, el cost en recursos humans creix desmesuradament en aquest model. L'equip de persones que treballa en un projecte basat en el model selectiu esmerça moltes hores en el

processos de selecció de recursos, gestió i acords amb les organitzacions productores i sobretot amb la descripció dels recursos.

- **Descontextualització de la col·lecció:** quan es fa la selecció de recursos la majoria de vegades no es té en compte el seu context, cosa que pot comportar el trencament de l'estructura d'enllaços, i per tant l'aparició de recursos orfes.

Annex 3. Llistat de projectes relacionats amb la preservació de recursos digitals

AOLA[99]

- Projecte desenvolupat per l'l'Österreichische National-bibliothek (Àustria).
- Iniciat el 1999, i s'ha paralitzat durant alguns períodes per manca de fons
- Model integral amb alguns elements del model híbrid
- Arxiu de 150 Gb., format per 2,7 milions de pàgines web de 21.000 llocs web (juny 2001)
- Creixement diari previst de 7 Gb.
- Inicialment funcionava amb el programa NEDLIB, però en una fase posterior passa a funcionar amb el programa COMBINE
- No consultable en línia. Accés limitat a investigadors, historiadors i estudiants

WebArchiv[100]

- Projecte desenvolupat per la Národní knihovna Ceské Republiky (República Txeca), juntament amb la Biblioteca de Moravia i l'Institute of Computer Science of Masaryk University
- Iniciat el 2000 amb el finançament del Ministeri de Cultura de la República Txeca
- Inicialment segueix el model integral, però va tendint cap a un model híbrid, incloent col·leccions temàtiques
- Arxiu de 2 Tb, format per 26 milions de pàgines (abril de 2006)
- Inicialment usava el programa NEDLIB, però en una següent fase ha implementat el programari de codi obert Heritrix, Nutch Wax i Wera
- Descripció dels recursos basat en les metadades Dublin Core

[Archiving the French web][101]

- Projecte liderat per la Biblioteca Nacional de França
- Iniciat el 2000
- Segueix el model híbrid
- L'abast del projecte inclou captures automàtiques a gran escala, captures sistemàtiques i contínues d'una selecció de llocs web (un 10% del total)
- També inclou alguna col·lecció temàtica com les eleccions franceses del 2002

---

[99] ÖSTERREICHISCHE NATIONAL-BIBLIOTHEK. *Austrian On Line Archive* [En línia]. Viena, Österreichische National-bibliothek, 2002. [Data de consulta: 15/11/2006]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola>.

[100] NÁRODNÍ KNIHOVNA CESKÉ REPUBLIKY. *WebArchiv: Archive of the web Czech* [En línia]. Praga: 2006. [Data de consulta 15/11/2006]. Disponible a: <http://en.webarchiv.cz/>

[101] MASANÉS, JULIEN. *Archiving the Web: experiments at the BnF* [En línia]. París: Bibliothèque Nationale de France, 2002. [Data de consulta: 15/11/2006]. Disponible a: <http://www.dpconline.org/graphics/events/presentations/pdf/Masanes.pdf>.

- Part dels recursos electrònics capturats es poden visualitzar a través del catàleg de la Biblioteca

[Archiving the Geek web][102]

- Liderat per l'Athens University of Economics and Business (Grècia).
- Iniciat el 2003
- Segueix el model integral
- Arxiu de 891 Mb, format per 300.000 recursos electrònics (2003)
- Ús d'un programari propi usat en altres projectes (SEWeP i Thesus)
- Sistema de captura basat en cerques a través d'algoritmes

Deposit.ddb.de[103]

- Liderat per Die Deutsche Bibliothek (Alemanya)
- Iniciat el 1997
- Model integral amb elements del model híbrid
- Catalogació per metadades
- Les proves inicials es van realitzar amb el web del govern alemany
- A partir de 2002 s'arriba a acords amb editors alemanys

E-Collection[104]

- Liderat per Libraries and Archives Canada
- Projecte iniciat el 1994
- Segueix el model selectiu
- E-Collection parteix del projecte EPPP (Electronic Publications Pilot Project) de 1994-95
- La seva finalitat inicial estava en la preservació de publicacions periòdiques electròniques a text complet
- A partir del 2004 es comencen a integrar tesis doctorals en línia i pàgines web

---

[102] ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS. Archiving the Greek Web [En línia]. Atenes: Departament d'Informàtica de la Universitat d'Atenes, 2003. [Data de consulta: 15/11/2006]. Disponible a: <http://www.iwaw.net/04/Lampos.pdf>.

[103] SIEGENTHALER, ANNETTE I EFFELSBERG, HANNELORE. Archive Server DEPOSIT.D-NB.DE [En línia]. Berlín: Deutsche National Bibliothek, juny de 2006. [Data de consulta: 15/11/2006]. Disponible a: <http://deposit.ddb.de/index_e.htm>.

[104] LIBRARIES AND ARCHIVES CANADA. *Electronic collection: a virtual collection of monographs and periodicals* [En línia]. Canadà: Libraries and Archives Canada, setembre 2006. [Data de consulta: 16/11/2006]. Disponible a: <http://epe.lac-bac.gc.ca/>.

e-Depot[105]

- Projecte liderat per la Koninklijike Bibliotheek (Països Baixos)
- Iniciat el 1995
- Segueix el model selectiu
- Bàsicament conté publicacions periòdiques electròniques científiques
- Funciona eminentment mitjançant acords amb editors
- És un projecte referent pel que fa a la preservació de recursos digitals
- Conté tres milions de números de revistes (març 2005)

EVA[106]

- Projecte liderat per la Helsingin yliopiston kirjasto (Finlàndia)
- Iniciat el 1997
- Segueix el model integral amb alguns aspectes del model híbrid
- La Biblioteca Nacional de Finlàndia lidera el Nordic Web Arxhive, que pretén la creació de l'arxiu web dels països escandinaus
- El projecte inicialment tenia com a objectiu les publicacions periòdiques electròniques, però actualment es dedica a la captura del web *.fi*
- La primera captura exhaustiva (març 1998) va recollir 1,800.000 pàgines web provinents de 7.500 diferents webs
- Avui en dia es calcula que el creixement de l'arxiu és de 500 Gb per any

New Zealand's digital heritage[107]

- Projecte liderat per la National Library of New Zealand
- Iniciat el 1999
- Segueix el model híbrid des dels seus inicis
- Fa captures temàtiques selectives relacionades amb esdeveniments significatius
- La llei de dipòsit legal novazelandesa de 2003 creava un escenari òptim per la preservació dels recursos en línia
- L'accés a l'arxiu web no és en obert
- Sota la motivació d'aquest projecte s'ha desenvolupat el programa *Web Curator*, una eina que facilita la validació i catalogació dels recursos digitals

---

[105] KONINKLIJIKE BIBLIOTHEEK. *e-Depot and digital preservation* [En línia]. Amsterdam: Koninklijike Bibliotheek, 2006. [Data de consulta: 16/11/2006]. Disponible a: <http://www.kb.nl/dnp/e-depot/e-depot-en.html>.

[106] LOUNAMAA, KIRSTI I SALONHARJU, INKERI. *EVA: the acquisition and archiving of electronic network publications in Finland* [Èn línia]. Helsinki: National Library of Finland, Center for Scientific Computing, 2003. [Data de consulta: 16/11/2006]. Disponible a: <http://www.ercim.org/publication/ws-proceedings/DELOS6 /eva.rtf.>.

[107] NATIONAL LIBRARY OF NEW ZELAND. *National Library to capture New Zealand's digital heritage* [En línia]. Wellington: National Library of New Zeland, 2004. [Data de consulta: 17/10/2006]. Disponible a: <http://www.natlib.govt.nz/en/contact/index.html>.

- El pressupost global del patrimoni digital és de 14 milions d'euros (2004)

Paradigma[108]

- Projecte liderat per la Nasjonalbiblioteket (Noruega)
- Iniciat el 2001
- Segueix el model integral amb elements del model híbrid
- Captures anuals, a partir de 2003 s'hi va incloure la captura dels dominis internacionals amb contingut noruec, així com 65 diaris digitals
- Noruega té des de 1990 una llei de dipòsit legal que empara els recursos digitals
- Desenvolupament del dipòsit DSM (Digital Storage Vault), amb capacitat per 100 Tb, habilitat només per lectura
- La indexació dels recursos es realitza de forma automàtica per mitjà dels programari FAST

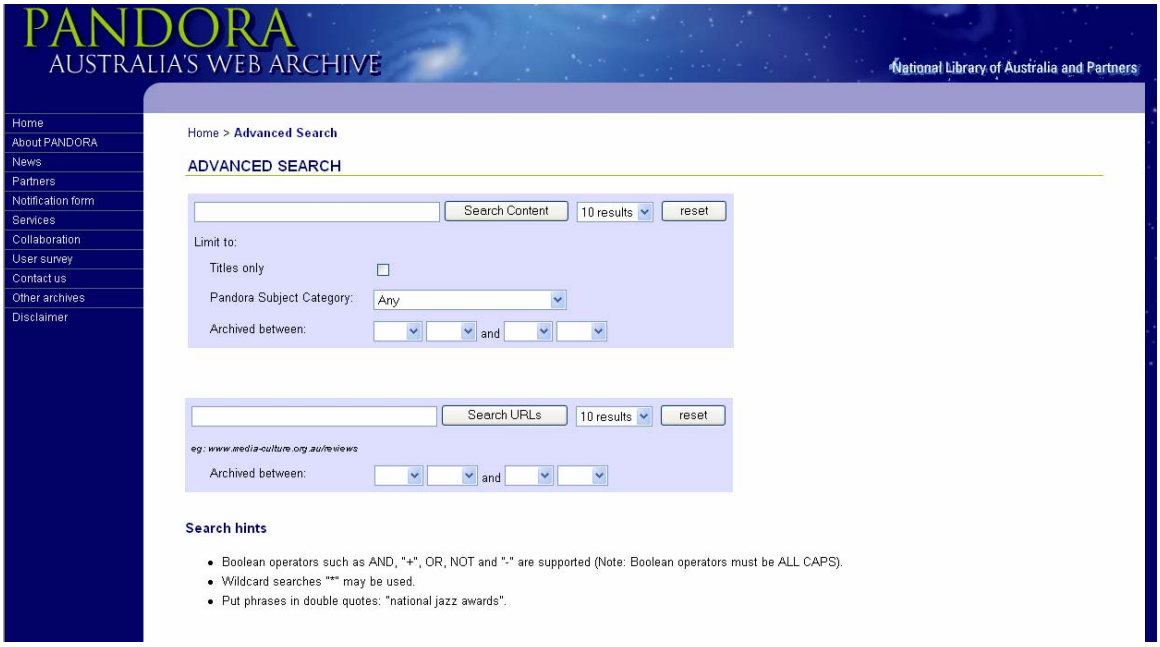
WARP[109]

- Projecte liderat per la National Diet Library (Japó)
- Iniciat el 2002
- Segueix el model selectiu
- La Biblioteca Nacional del Japó des del 2004 realitza captures per mitjà d'acords
- Actualment té capturats gairebé 2000 webs de l'administració pública, universitats, empreses, etc
- Es pot accedir a través de WARP a gairebé 1.500 números de revistes electròniques

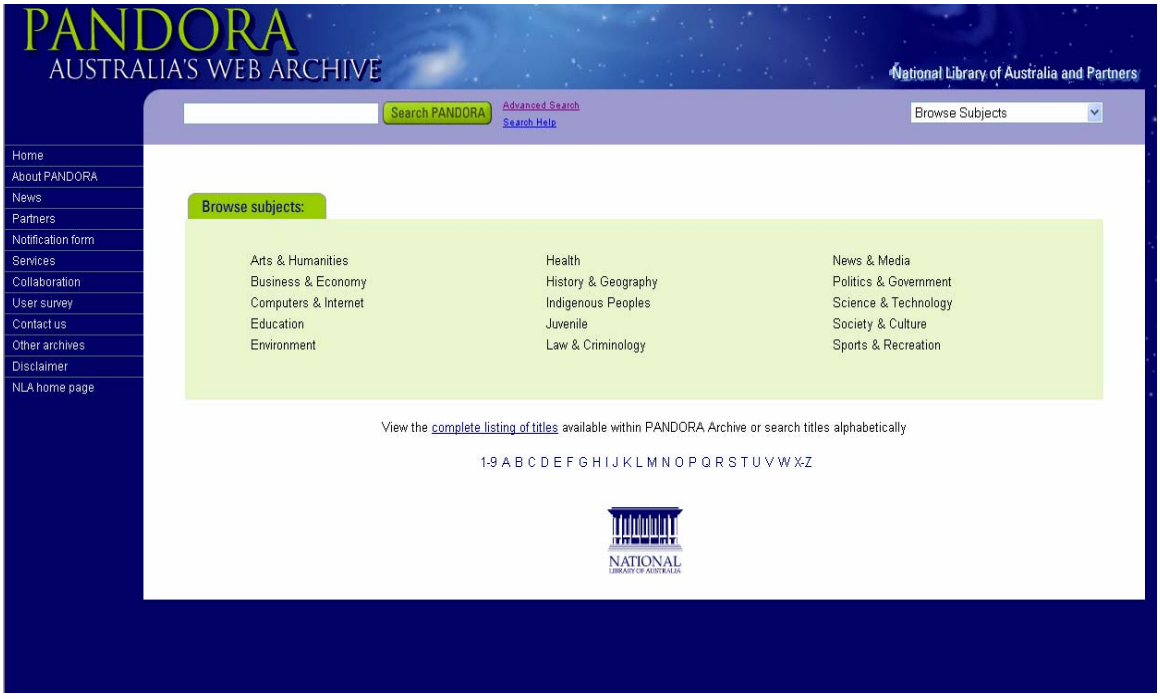
[108] ALBERTSEN, KETIL. *The Paradigma Web Harvesting Environment* [En línia]. Oslo: National Library of Norway, 2003. [Data de consulta: 16/10/2006]. Disponible a: <bibnum.bnf.fr/ECDL/2003/ proceedings.php?f=albertsen>.

[109] NATIONAL DIET LIBRARY. *WARP: web archiving prject* [En línia]. Japó: Digital Information Planning Office, juliol de 2006 [Data de consulta 16/10/2006]. Disponible a: <http://warp.ndl.go.jp /warp_eng.pdf>.

Annex 4. Sistemes de cerca del projecte PANDORA



Mòdul de cerca avançada

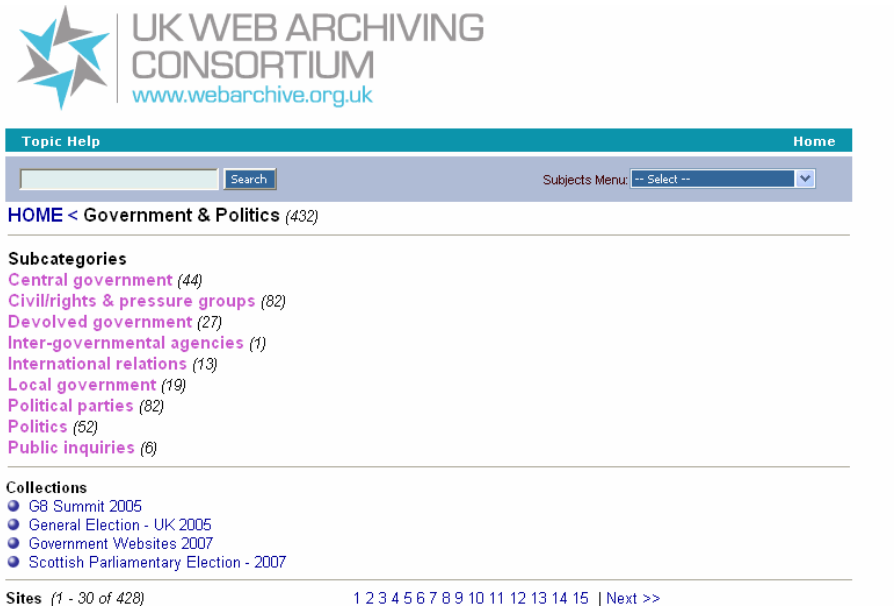


Mòdul de navegació per categories temàtiques

Annex 5. **Col·leccions temàtiques de la categoria de Política i Govern del Projecte PANDORA**

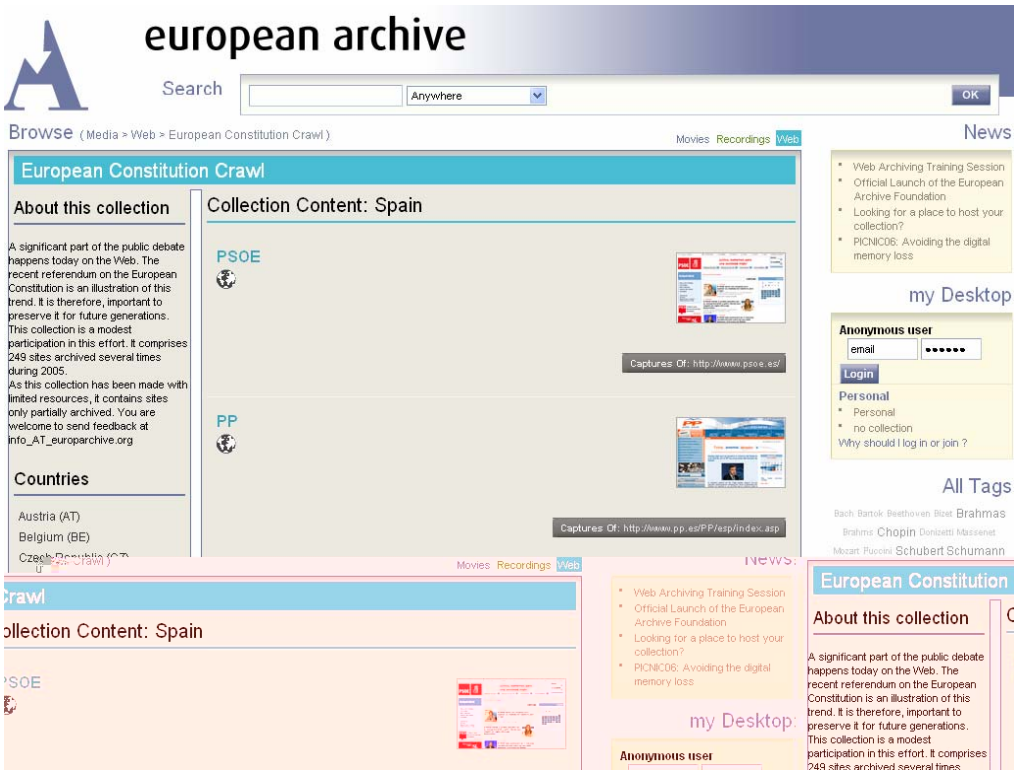
Annex 6. Projecte UK Web Archive

Mòdul de navegació per categories temàtiques

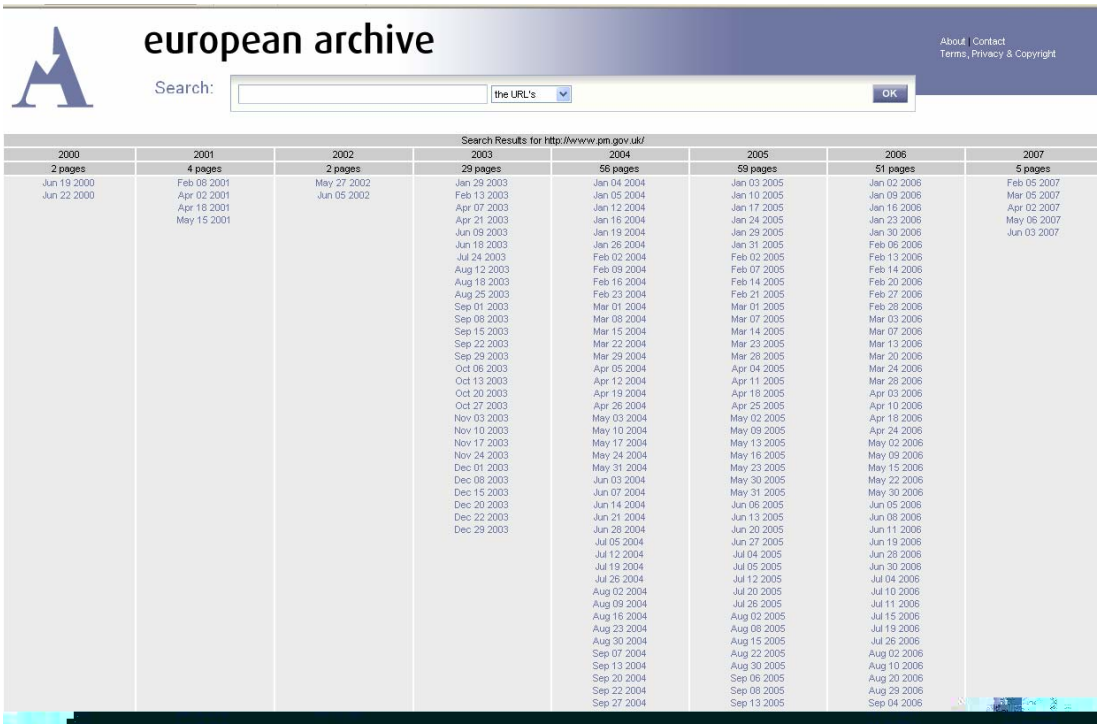


Col·leccions temàtiques de la categoria de Govern i Política

Annex 7. Projecte European Archive



Captures de webs de partits polítics espanyols en la Col·lecció temàtica de la Constitució Europea



Períodes de captura del 10 Downing Street Website

Annex 8. Projecte MINERVA

Sample MODS record

```
<mods>
  <titleInfo><title>FranUlmer.com -- Home Page</title></titleInfo>
  <titleInfo type="alternative"><title>Fran Ulmer, Democratic candidate for Governor,
    Alaska, 2002</title></titleInfo>
  <name type="personal"><namePart>Ulmer, Fran</namePart></name>
  <genre>Web site</genre>
  <originInfo>
    <dateCaptured point="start"encoding="iso8601">20020702
    </dateCaptured>
    <dateCaptured point="end" encoding="iso8601">20021203
    </dateCaptured>
  </originInfo>
  <language authority="iso639-2b">eng</language>
  <physicalDescription>
    <internetMediaType>text/html</internetMediaType>
    <internetMediaType>image/jpg</internetMediaType>
  </physicalDescription>
```

Registre catalogràfic basat en l'schema MODS

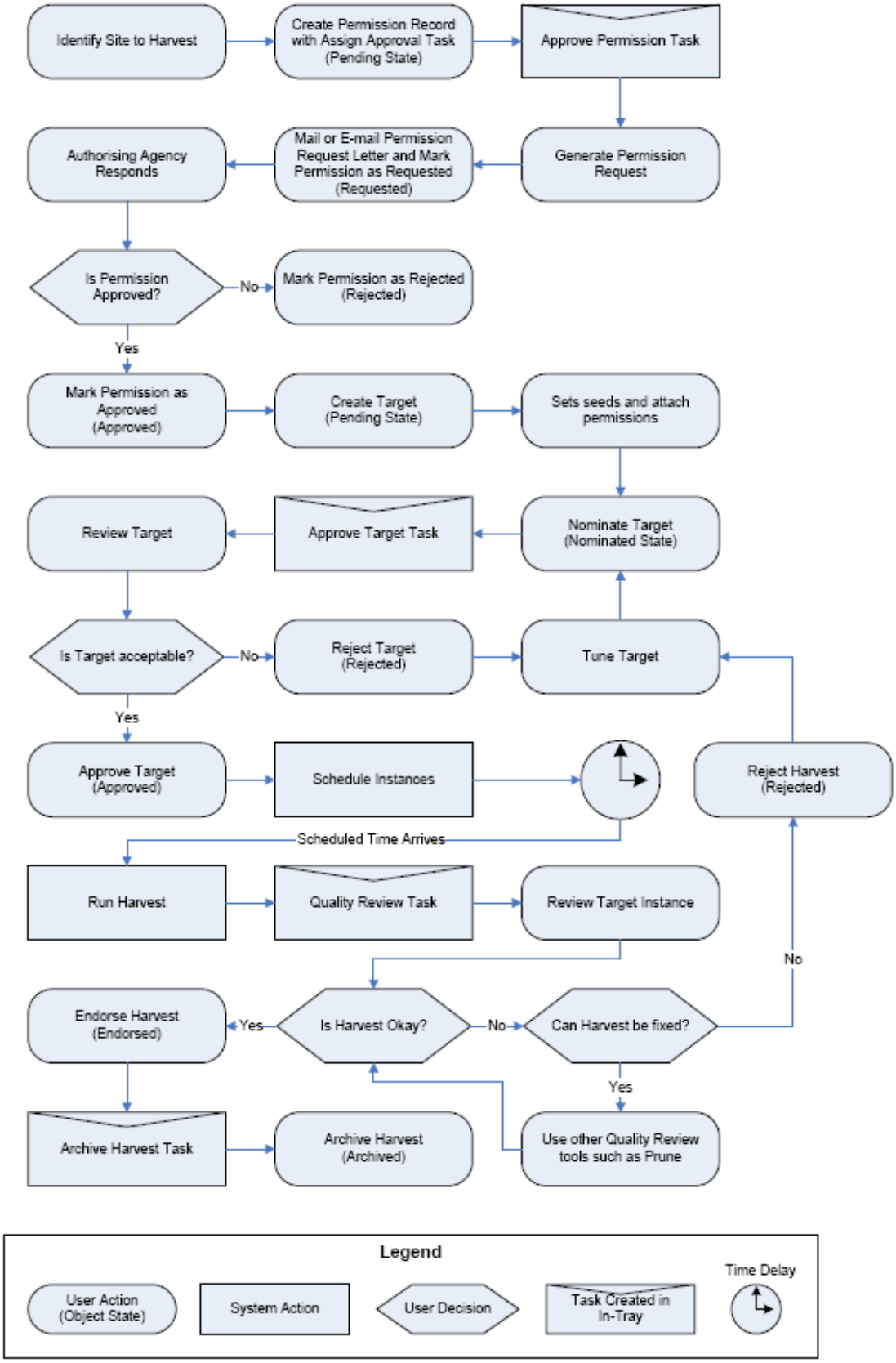
Descripció de la col·lecció de l'atac terrorista de l'11 de setembre de 2001

Annex 9. Projecte WebArchiv

Secció dedicada a les col·leccions temàtiques

Secció de cerca de recursos del dipòsit

Annex 10. Sistema de workflow de l'aplicació Web Curator



Bibliografia

Bibliografia general

AGUILLO, ISIDRO. *Internet Invisible* [En línia]*: los contenidos son la clave*. InternetLab : CINDOC-CSIC, abril 2003. [Data de consulta: 08/05/2006]. Disponible a: <http://internetlab.cindoc.csic.es /cursos/Internet_ Invisible2003.pdf>.

BIBLIOTECA DE CATALUNYA. *Llista d'encapçalaments de matèria en Català* [En línia]: LEMAC. Barcelona: Biblioteca de Catalunya, 2005. [Data de consulta: 02/05/2007]. Disponible a: <http://www.bnc.es/ catalegs/autoritats/lemac.php>.

BIBLIOTECA DE CATALUNYA. *Llista d'encapçalaments de noms i títols* [En línia]: LENOTI. Barcelona: Biblioteca de Catalunya, 2005. [Data de consulta: 02/05/2007]. Disponible a: <http://www.bnc.es/ catalegs/autoritats/lemac.php>.

COL·LEGI OFICIAL DE BIBLIOTECARIS I DOCUMENTALISTES DE CATALUNYA. *Quant cobra un bibliotecari-documentalista?* [En línia]. Barcelona, COBDC, abril de 2007. [Data de consulta 23/05/2007]. Disponible a: <http://www.cobdc.org/serveis/assessoria.html#01>.

ESNIC. *Registro de dominios ".es"* [En línia]. Madrid: Ministerio de Industria, Turismo y Comercio, 2006. [Data de consulta: 12/08/2006]. Disponible a: <https://www.nic.es/>.

GENERALITAT DE CATALUNYA. DEPARTAMENT DE GOVERNACIÓ I ADMINISTRACIONS PÚBLIQUES. *Calendari electoral de les eleccions municipals 2007* [En línia]. Barcelona: Generalitat de Catalunya, febrer de 2007. [Data de consulta: 23/03/2007]. Disponible a: <http://www.gencat.net/governacioap/eleccions/LOCALS 2007/MUNICIPALS07/m07_cale.pdf>.

GRUPO RH ASESORES. *Informe de remuneración España* 2006-2007 [Recurs electrònic]. Madrid: Ceinsa, 2006.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *ISBD (ER)* [En línia]: International Standard Bibliographic Description for Electronic Resources. França: IFLANET, juliol de 1999 [Data de consulta 06/06/2007]. Disponible a: <http://www.ifla.org /VII/s13/pubs/isbd.htm>.

IP2Location. *IP Address to Country, Region, City, Latitude, Longitude, ZIP Code, Internet Service Provider (ISP) and Domain Name* [En línia]. Bradenton: IP2Location, octubre de 2006. [Data de consulta: 12/08/2006]. Disponible a: <http://www.ip2location.biz/>.

LAMARCA, DOLORS. "La Biblioteca de Catalunya en el sistema bibliotecari de Catalunya [En línia]". *BiD: textos universitaris de biblioteconomia i documentació*, juny de 2004, núm. 12.

Barcelona: Facultat de Biblioteconomia i Documentació. [Data de consulta: 05/05/2006]. Disponible a: <http://www2.ub.es/ bid/consulta_articulos.php?fichero=12lamarc.htm>.

MAS I HERNÁNDEZ, JORDI. *La salut del català a Internet el 2005* [En línia]. Barcelona: Softcatalà, octubre de 2005. [Data de consulta 02/05/2007]. Disponible a: <http://www.softcatala.org /articles/article60.htm>.

MUNICIPIS I COMARQUES DE CATALUNYA. *Sistema d'informació d'administració local* [En línia]. Catalunya: MUNICAT, 2007. [Data de consulta: 22/03/2007]. Disponible a: <http://www.municat.net/pagines/ descarregues/desc.htm>.

PANTALEONI, ANA. "El dominio'.cat' obtiene 21.000 registros en un año" [En línia]. *El País.com Cataluña* Madrid, Prisacom S.A., gener de 2007. [Data de consulta: 02/05/2007]. Disponible a: <http://www.elpais.com/articulo/cataluna/dominio/cat/obtiene/21000/registros/ano/elpepuespcat/ 20070215elpcat_20/Tes>.

PARTAL, VICENT. *El català a la xarxa* [En línia]*: història i raons d'un cas d'èxit*. Barcelona: Softcatalà, abril de 2004. [Data de consulta: 02/05/2007]. Disponible a: <http://www.softcatala.org/articles/article39.htm>.

RIPE. *Network coordination center* [En línia]. Amsterdam: RIPE, 2006. [Data de consulta: 12/08/2006]. Disponible a: <http://www.ripe.net/>.

SOLER MARTÍ, JOSEP. *Balanç de l'ús del català a Internet al 2006* [En línia]. Barcelona: Softcatalà, gener de 2007. [Data de consulta: 02/05/2007]. Disponible a: <http://www.softcatala.org/noticies/03012007510.htm>.

WIKISOURCE. *Llei de Biblioteques de Catalunya, de 24 d'abril de 1981* [En línia]. Desembre 2005. [Data de consulta: 10/06/2006]. Disponible a: <http://ca.wikisource.org/wiki /Llei_de_biblioteques_de_Catalunya_1981>.

PAREJA, VÍCTOR MANUEL; ORTEGA, JOSÉ LUÍS; 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.

Bibliografia PADICAT

BIBLIOTECA DE CATALUNYA. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)* [En línia]. Barcelona: Biblioteca de Catalunya, desembre 2005. [Data de consulta: 09/05/2007]. Disponible a: <http://www.recercat.net/handle/2072/1757>.

BIBLIOTECA NACIONAL DE CATALUNYA. *PADICAT*: *Patrimoni Digital de Catalunya* [En línia]. Barcelona: Biblioteca Nacional de Catalunya, 2006. [Data de consulta: 24/10/2006]. Disponible a: <http://www.padicat.cat/quees.php>.

LLUECA, CIRO. "Archivando la web, el proyecto PADICAT (Patrimonio Digital de Catalunya)". *El profesional de la información*, vol. 15, núm 6, p. 473-478. Disponible a: <http://eprints.rclis.org/archive/00007767/01/epi_padicat.pdf>.

LLUECA, CIRO. "El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya". *10es Jornades Catalanes d'Informació i Documentació*. Barcelona, maig 2006. <http://eprints.rclis.org/archive/00006434/01/llueca_padicat.pdf>.

SERRA, EUGÈNIA. "Archivando la Web catalana: iniciativas cooperativas de preservación digital en Catalunya". *La Recuperación de la memoria, muchas más oportunidades que realidades: el trabajo cooperativo de archivos, bibliotecas y museos*. Universidad del País Vasco, 23 a 25 d'agost de 2006. Disponible a: <http://www.bnc.es/bc/archivando_web_catalana.pdf>.

Bibliografia tecnològica

ASCHENBRENNER, ANDREAS. "AOLA: the austrian on-line archive". *Long-term preservation of digital material -building an archive to preserve digital cultural heritage from the internet* [En línia]. Viena: Information & SoftwareEngineering Group, 2004. [Data de consulta: 26/01/2007]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/AOLA.html>.

ASCHENBRENNER, ANDREAS. "Adapting the Combine crawler". *Long-term preservation of digital material -building an archive to preserve digital cultural heritage from the internet* [En línia]. Viena: Information & SoftwareEngineering Group, 2004. [Data de consulta: 26/01/2007]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Adapting_Combine_crawler.html>.

BIBLIOTECA NACIONAL D'AUSTRÀLIA. Pandora Digital Archiving System (PANDAS) [En línia]. Canberra: Biblioteca Nacional d'Austràlia, març de 2006. [Data de consulta 01/02/2006]. Disponible a: < http://pandora.nla.gov.au/pandas.html/>.

HAKALA, JUHA. "The Nedlib harvester" [En línia].*NEDLIB workshop 2000*. La Haya: desembre del 2000. [Data de consulta: 26/01/2007]. Disponible a: <http://nedlib.kb.nl/workshop/ NEDLIB%20harvester.ppt#1>.

IBM. *IBM Servers* [En línia]. New York: IBM, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://www-03.ibm.com/servers/>.

INTERNACIONAL INTERNET PRESERVATION CONSORTIUM; NORDIC WEB ARCHIVE. *Wera* [En línia]. Internet Archive, 2005. [Data de consulta 04/05/2007]. Disponible a: <http://archive-access.sourceforge.net/ projects/wera/>.

INTERNET ARCHIVE, ET AL. *Heritrix User Manual* [En línia]. Internet Archive, 2005. [Data de consulta: 23/01/2007]. Disponible a: <http://crawler.archive.org/articles/user_manual/ index.html>.

KNOWLIB. *The Combine harvesting robot* [En línia]. Lund**:** Department of Information Technology, Lund University, gener de 2007. [Data de consulta: 26/01/2007]. Disponible a: <http://combine.it.lth.se/ #features>.

NOTESS, GREG R. "The Wayback Machine [En línia]: the Web's Archive". *Info Today*, vol 26, núm. 2, març - abril de 2002. [Data de consulta: 04/05/2007]. Disponible a: <http://www.infotoday.com/ online/mar02/OnTheNet.htm>.

PRICOINSA. *Bienvenido a Pricoinsa* [En línia]. Barcelona: Pricoinsa, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://www.pricoinsa.es/>.

NATIONAL LIBRARY OF NEW ZEALAND; BRITISH LIBRARY. *Web Curator Tool Quick Start Guide*[En línia]. National Library of New Zealand & British Library, setembre de 2006. [Data de consulta: 03/04/2007]. Disponible a: <http://webcurator.sourceforge.net/docs/1.1/wct-1.1-quick-start-guide.pdf>.

ROCHE, XAVIER, ET AL. *HTTrack Web Copier* [En línia]. França: Leto Kauler, 2007. [Data de consulta: 26/01/2007]. Disponible a: <http://www.httrack.com/page/1/en/index.html>.

SIGUROSSON, KRISTINN. *Managing duplicates across sequential crawls* [En línia]. Reykjavík: National and University Library of Iceland, 2007. [Data de consulta: 05/06/2007]. Disponible a: <http://vefsofnun.bok.hi.is/upload/3/ManagingDuplicatesAcrossSequentialCrawls.pdf>.

SUN MICROSYSTEMS. *Productos y servicios* [En línia]. Madrid: Sunmicrosystems, 2007. [Data de consulta: 25/05/2007]. Disponible a: <http://es.sun.com/productos_servicios/>.

Bibliografia sobre preservació

BIBLIOTECA NACIONAL D'AUSTRÀLIA. *Arching web resources* [En línia]. Canberra: Biblioteca Nacional d'Austràlia, novembre 2004. [Data de consulta 13/10/2006]. Disponible a: <http://www.nla.gov.au/ webarchiving/>.

BIBLIOTECA NACIONAL D'AUSTRÀLIA. *Directrices para la preservación del patrimonio digital* [En línia]. Canberra: Unesco, 2003. [Data de consulta: 05/05/2006]. Disponible a: <http://unesdoc.unesco.org /images/0013/001300/130071s.pdf>.

BIBLIOTECA NACIONAL D'AUSTRÀLIA. *Preserving Access to Digital Information* [En línia]. Canberra: Biblioteca Nacional d'Austràlia, 2006. [Data de consulta: 13/10/2006]. Disponible a: <http://www.nla.gov.au/ padi/index.html>.

CORDÓN, JOSÉ ANTONIO. "El depósito legal y los recursos digitales en línea [En línia]". A: *Las bibliotecas nacionales del siglo XXI*. Valencia: Biblioteca Valenciana, 2005. [Data de consulta: 06/05/2006]. Disponible a: <http://bv.gva.es/documentos/Ponencias/Cordon.pdf>.

HAKALA, JUHA. *Archiving the web* [En línia]: *European experiences*. Helsinki: Helsinki University Library, octubre de 2003. [Data de consulta: 03/02/2007]. Disponible a: <http://www.lib.helsinki.fi/tietolinja/0203/webarchive.html>.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. *Netpreserve.org* [En línia]. França: National Library of France, setembre de 2006. [Data de consulta 12/10/2006]. Disponible a: <http://www.netpreserve.org/ about/index.php>.

IWAW. *6th International* W*eb Archiving Workshop* [En línia]. 2006. [Data de consulta: 12/12/2006]. Disponible a: <http://www.iwaw.net/06/>.

LLUECA FONOLLOSA, CIRO. "Webs sempre accessibles [En línia]: les biblioteques nacionals i els dipòsits digitals nacionals". *BiD: textos universitaris de biblioteconomia i documentació*, desembre 2005, núm. 15. Barcelona: Facultat de Biblioteconomia i Documentació. [Data de consulta: 05/05/2006]. Disponible a: http://www2.ub.edu/bid/consulta_ articulos.php?fichero= 15lluec1. htm>.

MINERVA PROJECT. *107th Congress Web Archive* [En línia]. Washington: Library of Congress, octubre de 2006. [Data de consulta: 15/01/2007]. Disponible a: <http://lcweb2.loc.gov/cocoon/minerva/html /107th/ search.html>.

MINERVA PROJECT. *September 11 Web Archive* [En línia]. Washington: Library of Congress, octubre de 2006. [Data de consulta: 15/01/2007]. Disponible a: <http://lcweb2.loc.gov/cocoon/minerva/html/sept11 /sept11-about.html>.

Nordic Web Archive. *NWA* [En línia]. Noruega: 2006. [Data de consulta: 12/10/2006]. Disponible a: <http://nwa.nb.no/>.

Bibliografia sobre projectes

ANDERSEN, BJARNE. *The DK-domain: in words and figures* [En línia]. Uhus: State & University Library, 2006. [Data de consulta: 28/11/2006]. Disponible a: <netarchive.dk/publikationer/ DFrevy_english.pdf>.

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS. *Archiving the Greek Web* [En línia]. Atenes: Departament d'Informàtica de la Universitat d'Atenes, 2003. [Data de consulta: 15/11/2006]. Disponible a: <http://www.iwaw.net/04/Lampos.pdf>.

CHRISTENSEN-DALSGAARD, BIRTE. "Web Archive Activities in Denmark" [En línia]. *RLG Diginews*. Califòrnia: RLG, Vol. 3 núm. 8., 15 de juny de 2004. [Data de consulta: 28/11/2006]. Disponible a: <http://www.rlg.org/en/page.php?Page_ID=17661#article0>.

EUROPEAN ARCHIVE FOUNDATION. *European Archive* [En línia]. [Amsterdam], European Archive Foundation, 2006. [Data de consulta: 05/02/2007]. Disponible a: <http://www.europarchive.org /index.php>.

INTERNET ARCHIVE. *Internet Archive* [En línia]. San Francisco: 2006. [Data de consulta: 12-10-2006]. Disponible a: <http://www.archive.org/about/about.php>.

KONINKLIJIKE BIBLIOTHEEK. *e-Depot and digital preservation* [En línia]. Amsterdam: Koninklijike Bibliotheek, 2006. [Data de consulta: 16/11/2006]. Disponible a: <http://www.kb.nl/dnp/e-depot/e-depot-en.html>.

LIBRARIES AND ARCHIVES CANADA. *Electronic collection: a virtual collection of monographs and periodicals* [En línia]. Canadà: Libraries and Archives Canada, setembre 2006. [Data de consulta: 16/11/2006]. Disponible a: <http://epe.lac-bac.gc.ca/>.

LIBRARY OF CONGRESS. *Web Capture & Archiving* [En línia]. Washington: Library of Congress, abril de 2003. [Data de consulta: 06/02/2007]. Disponible a: <http://www.loc.gov/acq/devpol/ webarchive.html>.

LOUNAMAA, KIRSTI I SALONHARJU, INKERI. *EVA: the acquisition and archiving of electronic network publications in Finland* [Èn línia]. Helsinki: National Library of Finland, Center for Scientific Computing, 2003. [Data de consulta: 16/11/2006]]. Disponible a: <http://www.ercim.org/publication/ws-proceedings/DELOS6 /eva.rtf.>.

MASANES, JULIEN. *Archiving the Web: experiments at the BnF* [En línia]. París: Bibliothèque Nationale de France, 2002. [Data de consulta: 15/11/2006]. Disponible a: <http://www.dpconline.org/graphics/events/ presentations/pdf/Masanes.pdf>.

Národní knihovna Ceské Republiky. *WebArchiv: Archive of the web Czech* [En línia]. Praga: 2006. [Data de consulta 15/11/2006]. Disponible a: <http://en.webarchiv.cz/>.

NATIONAL DIET LIBRARY. *WARP: web archiving prject* [En línia]. Japó: Digital Information Planning Office, juliol de 2006 [Data de consulta 16/10/2006]. Disponible a: <http://warp.ndl.go.jp /warp_eng.pdf>.

NATIONAL LIBRARY OF NEW ZELAND. *National Library to capture New Zealand's digital heritage* [En línia]. Wellington: National Library of New Zeland, 2004. [Data de consulta: 17/10/2006]. Disponible a: <http://www.natlib.govt.nz/en/contact/index.html>.

NATIONAL LIBRARY OF SWEDEN. *Kulturarw3* [En línia]. National Library of Sweden, març de 2007. [Data de consulta: 26/04/2007]. Disponible a: <http://www.kb.se/kw3/ENG/>.

ÖSTERREICHISCHE NATIONAL-BIBLIOTHEK. *Austrian On Line Archive* [En línia]. Viena, Österreichische National-bibliothek, 2002. [Data de consulta: 15/11/2006]. Disponible a: <http://www.ifs.tuwien.ac.at/~aola>.

PERSSON, KRISTER; ET AL. *The Kulturarw*[3] Project [En línia] : *the Royal Swedish Web Archiw e*[3]. Estocolm: National Library of Sweden, [200?]. [Data de consulta: 25/04/2007]. Disponible a: <www.ifla.org/IV/ifla66/papers/154-157e.htm>.

SIEGENTHALER, ANNETTE I EFFELSBERG, HANNELORE. Archive *Server DEPOSIT.D-NB.DE* [En línia]. Berlín: Deutsche National Bibliothek, juny de 2006. [Data de consulta: 15/11/2006]. Disponible a: <http://deposit.ddb.de/index_e.htm>.

UKWAC. *G8 summit 2005* [En línia]: *related Internet sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col/c8150.html>.

UKWAC. *General Election - UK 2005* [En línia]: *related Internet Sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col/c8100.html>.

UKWAC. *Terrorist attack - London, 7th July 2005* [En línia]: *related Internet Sites*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/col /c8125.html>.

UKWAC. *UK Web Archiving Consortium* [En línia]: *Archive*.United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://www.webarchive.org.uk/>.

UKWAC. *UK Web Sites Maintained only in the UKWAC Archive* [En línia]: *Report Number 2 September 2006*. United Kingdom: UKWAC, 2006. [Data de consulta: 02/02/2007]. Disponible a: <http://info.webarchive.org.uk/mia_reports/ukwac_mia_list_sept_2006.pdf>.

Bibliografia d'estratègies de captura

ALBERSTSEN, KETIL. *The Paradigma Web Harvesting Environment* [En línia]. Oslo: National Library of Norway, 2003. [Data de consulta: 16/10/2006]. Disponible a: <bibnum.bnf.fr/ECDL/2003/ proceedings.php?f=albertsen>.

BAEZA-YATES, R. A.; CASTILLO, C. "Crawling the infinite Web: Five levels are enough". *Workshop on algorithms and models far the Web-Graph WAW2004*. Roma: Springer Verlag, 2004.

DRAKOS, NIKOS i MOORE, ROSS. *Documentation for the Combine (focused) crawling system* [En línia]. Traducció de Anders Ardö. Lund: Department of Information Technology, Lund University, febrer de 2007. [Data de consulta: 01/03/2007]. Disponible a: <http://combine.it.lth.se/documentation/>.

ESTER, MARTIN; GROß, MATTHIAS; KRIEGEL, HANS-PETER. "Focused Web Crawling: [En línia] a generic framework for specifying the user interest and for adaptive crawling strategies". *Twenty-Seventh International Conference on Very Large Databases*, 2001. [Data de consulta: 25/01/2007]. Disponible a: <http://citeseer.ist.psu.edu/ester01focused.html>.

KOERBIN, PAUL. *Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005* [En línia]. Canberra: Biblioteca Nacional d'Australia, octubre de 2005. [Data de consulta: 27/01/2007]. Disponible a: <www.pandora.nla.gov.au/documents /domain_harvest_report_public.pdf>.

MARILL, J.; BOYKO, A.; ASHENFELDER, M.; ET AL. *Web Harvesting Survey* [En línia]. Washington: International Internet Preservation Consortium, 2005. [Data de consulta: 25/01/2007]. Disponible a: <http://www.netpreserve.org/publications/iipc-r-001.pdf>.

MASANES, JULIÉN. "Web archiving methods and approaches [En línia]: a comparative study". *Library Trends*, Vol. 54, núm. 1, juny de 2005. [Data de consulta: 25/01/2007]. Disponible a: Factiva.

NOVAK, BLAZ*. A survey of focused web crawling algorithms* [En línia]. Ljubljana: Department of Knowledge Technologies, Jozef Stefan Institute, 2005. [Data de consulta: 02/03/2007]. Disponible a: <eprints.pascal-network.org/archive/00000738/01/BlazNovak-FocusedCrawling.pdf>

SRINIVASAN, PADMINI; MENCZER, FILIPPO; PANT, GAUTAM. *Defining Evaluation Methodologies for Topical Crawlers* [En línia]. Iowa: The University of Iowa, 2004. [Data de consulta: 02/03/2007]. Disponible a: <dollar.biz.uiowa.edu/~pant/Papers/crawl_framework_position.pdf>.

ROMERO TRUJILLO, RAFAEL. Simulation *tool to study focused web crawling strategies* [En línia]. Lund: Department of Information Technology. Lund University, març de 2006. [Data de consulta: 03/03/2007]. Disponible a: <combine.it.lth.se/CrawlSim/>.

MEDELYAN, OLENA, Et. al. *Language Specific and Topic FocusedWeb Crawling* [En línia]. Freiburg: Jena: Freiburg University, Jena University, 2005. [Data de consulta: 03/03/2007]. Disponible a: <www.cs.waikato.ac.nz/~olena/publications/lrec2006_focused_crawler.pdf>.