bid
textos universitaris de biblioteconomia i documentació
número 15
desembre de 2005
ISSN 1575-5886
DL B-19.675-1998
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Webs sempre accessibles: les biblioteques nacionals i els dipòsits digitals nacionals

[Versión castellana]

CIRO LLUECA FONOLLOSA
Projecte PADICAT (Patrimoni Digital de Catalunya)
Biblioteca de Catalunya
cllueca@bnc.es

Resum [Abstract] [Resumen]

Les tecnologies de la informació i comunicació han facilitat que el patrimoni cultural i científic i la informació en general es presentin en format digital, com també en els formats analògics tradicionals. La reacció no s'ha fet esperar, i des de la dècada dels noranta han sorgit diversos projectes destinats a garantir l'accés permanent a la producció digital —la recopilació i l'emmagatzematge, el tractament, la preservació i la difusió. En aquest article es presenta la panoràmica mundial dels models existents de *dipòsits digitals nacionals*, nom que reben aquests projectes impulsats habitualment per les biblioteques nacionals amb un objectiu comú: fer que les pàgines web siguin sempre accessibles.

1 Introducció

Les tecnologies de la informació i la comunicació han facilitat que el patrimoni cultural i científic i la informació en general es presentin en format digital, com també en els formats analògics tradicionals. Actualment, i tal com ho exposen les *Directrices para la preservación del patrimonio digital*,[1] els recursos que són fruit del coneixement o l'expressió dels éssers humans, ja siguin de caràcter cultural, educatiu, científic o administratiu, o comprenguin informació tècnica, jurídica, mèdica i d'un altre tipus, es generen directament en format digital o es converteixen a aquest format a partir de material analògic ja existent. Els productes que d'antuvi es generen digitalment no existeixen en un altre format que no sigui l'electrònic original.

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

Les dificultats són notables. Per començar, els mètodes tradicionals de preservació de la producció bibliogràfica (com ara el dipòsit legal) són de difícil aplicació en l'entorn digital perquè, a banda de la possible obsolescència del text legal (el cas espanyol), els recursos digitals poden instal·lar-se en servidors d'arreu del món (com passa també amb els impressors que no són editors). Aquest fet dificulta la tradicional correspondència geogràfica entre la ubicació del productor i la llengua o la temàtica publicada. En segon lloc, la producció digital té un creixement exponencial i, a més, és molt variable la durabilitat dels materials publicats a Internet[2] i, en conseqüència, limitada la possibilitat d'accés permanent al patrimoni. Finalment, cal assenyalar la qüestió de la propietat intel·lectual del producte digital, que està mancat d'un dret basat en el principi de còpia per a la preservació que asseguri la conservació i perdurabilitat del patrimoni digital, amb les limitacions comercials que siguin necessàries.

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

El juny de 2005, la Biblioteca de Catalunya va posar en marxa el projecte PADICAT (Patrimoni Digital de Catalunya). El text que ara es presenta és un dels fruits de la primera fase del projecte. L'objectiu d'aquest article és presentar la panoràmica mundial en matèria d'accessibilitat permanent a la producció web,[3] i els models existents a l'hora de capturar les produccions digitals en línia. No s'examinen altres qüestions d'aquests models, com ara els aspectes legals de cada territori o les diferències dels programaris emprats per les biblioteques.

2 Dipòsits digitals nacionals: *arxivant la web*

En els orígens de la preservació digital podem esmentar les accions de les biblioteques virtuals, amb projectes de biblioteques de recerca, universitàries, nacionals i també públiques dedicats a presentar directoris temàtics de recursos electrònics. Complementàriament i per donar resposta a necessitats temàtiques concretes (normalment d'àmbit geogràfic que seguien el model enciclopèdic digital dels CD-ROM), es va optar per crear dipòsits multiformat (imatges, so, text, gràfics, etc.). El pas lògic següent ha estat conservar els recursos propis per garantir-ne l'accés amb totes les variables formals que s'han produït en el temps; són les iniciatives que es coneixen amb els noms de *dipòsits institucionals*, *arxius d'e-prints*, etc. Quan el procés es dedica a un territori, parlem dels *dipòsits*[4] *digitals nacionals*, dels *arxius web* o de les *biblioteques nacionals digitals*.

Un dipòsit digital nacional té la missió de garantir l'accés a llarg termini als recursos digitals que es generen en un territori, o sobre un territori determinat. De fet, la missió de la Biblioteca de Catalunya, com a biblioteca nacional que és, no és altra que recollir, conservar i difondre la producció bibliogràfica catalana i la relacionada amb l'àmbit lingüístic català, i vetllar per la conservació i la difusió del patrimoni bibliogràfic. I s'entén que aquest patrimoni bibliogràfic inclou també la producció bibliogràfica digital catalana que conformarà el PADICAT, el Patrimoni Digital de Catalunya.

2.1 Els models existents

Les experiències existents de dipòsits nacionals digitals s'agrupen en dos models inicials. Per una banda, el model integral o exhaustiu (model majoritari, i característic de Suècia, Noruega, Finlàndia, Islàndia i Àustria, entre d'altres) aposta per la integració automàtica del total de la web objecte de preservació a partir de determinats criteris infraestructurals (lingüístics, segons el domini de les web, segons la ubicació del servidor, etc.). I per l'altra, el model selectiu (assimilat per Austràlia, el Canadà, el Japó i el Regne Unit, entre altres països) dirigeix les accions de recopilació d'acord amb una política selectiva temàtica (sobre un espai geogràfic determinat, al voltant d'un tema d'interès nacional, etc.); per fer-ho arriba a acords amb els editors o productors de recursos web.

Aquests dos models han deixat pas en alguns països, però cada vegada amb més força, a models híbrids que complementen la captura periòdica del total de la web nacional amb acords amb els productors, que es basen en interessos temàtics o que tenen relació amb esdeveniments d'actualitat (eleccions, catàstrofes, etc.).

Finalment, altres anàlisis teòriques[5] de la situació apunten cap a una classificació més complexa (segons que la web que s'hagi de capturar sigui estàtica o dinàmica, per

exemple), però entenem aquí que la captura restrictiva de la web segons la seva complexitat de preservació i garantia d'accés (si és estàtica, és més fàcil capturar-la i preservar-la) és només un primer pas per a l'objectiu final de tots els dipòsits digitals nacionals.

2.1.1 El model integral

Els principals avantatges o punts forts del model integral són:

- *Riquesa de la col·lecció, en quantitat i en qualitat*. D'acord amb aquest model, no es condiciona selectivament què és interessant i què no ho és, sobretot si es considera que difícilment som capaços de preveure quins seran els usos i les línies d'investigació futures. En reflectir, a més, el creixement de la web i els canvis en el disseny de la publicació web de tots els nivells, aporta un component sociològic afegit, ja que les pàgines personals, els *weblogs*, els xats i *in extremis* els videojocs en línia formen part també de la producció digital nacional en els dipòsits integrals. Lligat a aquest punt fort hi ha el fet que la captura exhaustiva permet respectar en bona mesura la interrelació dels llocs web.

  A tall d'exemple, el projecte suec Kulturarw3 conté els llocs web amb els dominis .se i .nu,[6] els llocs web amb dominis internacionals (.com, .org i .net) ubicats en servidors en territori suec[7] i la *Suecana extreana* —les webs que parlen sobre Suècia, de viatges per Suècia o de traduccions d'obres literàries sueques.

- *Compilació automàtica*. El maquinari i programari emprats asseguren una capacitat alta de captura i d'emmagatzematge en resposta a uns paràmetres determinats, que poden ser tan amplis com la direcció del projecte estableixi. Aquest fet minimitza els recursos més costosos, els de personal, i alhora permet aconseguir resultats visibles (resultats presentables, "vendibles" a la sensibilitat política i ciutadana) en un període de temps raonablement curt, ja que en poc temps es crea una col·lecció significativa.

  Finalitzada la primera captura del projecte finlandès (agost–setembre de 2001), els responsables de la Helsingin Yliopiston Kirjasto-Suomen Kansalliskirjasto asseguren que s'han capturat 7,5 milions d'URL, i calculen que aquesta fotografia del web finlandès representa entre el 30 i el 50 % del total capturable.

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

  Mentre que el projecte australià Pandora (selectiu) dedica un total de tretze persones a temps complet i el sistema emprat al Quebec (selectiu, i integrat al catàleg IRIS) té un equip de deu persones, els projectes basats en la captura integral tenen equips sensiblement menors —per exemple, els projectes d'Àustria i Suècia tenen equips d'una a tres persones.

Els principals inconvenients o punts febles del model integral són:

- *Impossibilitat d'accedir a la Internet invisible*. El sistema automàtic de captura només té accés als recursos que són publicats en règim obert i, per tant, es produeixen llacunes en la captura de webs de pagament, webs protegides amb contrasenyes, pàgines òrfenes i la major part de les pàgines dinàmiques. Així mateix, tampoc no és possible accedir a les infranets o bases de dades bibliogràfiques (catàlegs de biblioteques) o alfanumèriques (diccionaris).

El fet és que, com han explicat repetidament Isidro Aguillo i els investigadors d'InternetLab,[8] la Internet invisible multiplica la Internet visible entre dues i cinquanta vegades. A més, aquest espai representa un arxipèlag de qualitat pel tipus de recursos que conté (articles, estudis científics, publicacions digitals, etc.).

- *Compilació irregular de la col·lecció*. Aquesta característica és conseqüència de les llacunes en el control dels ítems col·leccionats (per exemple, en les publicacions periòdiques), la no-reclamació dels documents no accessibles, la pèrdua de documents importants i dels canvis freqüents que es produeixen en determinats llocs web.

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

- *Accés limitat als resultats*. La manca d'un procés de catalogació (metadades, inclusió al catàleg de la biblioteca, etc.) dificulta la recuperació dels documents capturats. D'altra banda, el respecte dels drets d'autor per publicar sense autorització o sense acord previ els recursos capturats fa que es restringeixi l'accés als dipòsits nacionals i que normalment només es pugin consultar des de les mateixes instal·lacions de les biblioteques nacionals. Aquesta mesura no acaba de conjugar amb la pròpia naturalesa dels projectes: garantir l'accés a la producció digital.

  Dels projectes amb intenció exhaustiva, únicament Internet Archive ofereix accés obert i en línia als seus fons, i actualment només permet la cerca per URL. La resta de projectes —inclosos els escandinaus, que són els més veterans— limiten la consulta de les seves col·leccions a les dependències de les biblioteques nacionals que els lideren. Només una part molt petita de la col·lecció rep un tractament que facilita la recuperació més enllà de l'URL o la data de captura.

2.1.2 El model selectiu

Els principals avantatges o punts forts del model selectiu són:

- *Creació d'una col·lecció equilibrada*. Cada ítem que formarà part de l'arxiu és avaluat tenint en compte la seva pauta de publicació. El model respon, doncs, a pautes més properes al model bibliotecari clàssic, en el sentit que es coneix allò que forma part de la col·lecció i que aquesta s'amplia tenint en compte la realitat del territori i de tots els usuaris.

  Per exemple, el model de creixement de Pandora és visible també al Regne Unit, amb l'UKWA (*United Kingdom Web Archive*). El recurs presenta una classificació dels continguts molt similar a la del popular directori *Yahoo*. A partir d'un màxim de nou categories inicials (art i humanitats, negocis i economia, etc.), es produeix una taxonomia en cascada (art i humanitats: arquitectura; dansa; belles arts; geografia; història; llengües; literatura; música, etc.) i la presència dels recursos digitals britànics és temàticament equilibrada en totes i cadascuna.

- *Màxima facilitat d'accés al fons*. Cada ítem por ser completament catalogat i passar a formar part de la bibliografia nacional, de manera que les dades bibliogràfiques poden ser compartides. Els recursos s'integren en el catàleg de la biblioteca, i els acords permeten publicar els recursos en línia, obertament. La catalogació dels

documents fa que les possibilitats de recuperació siguin il·limitades.

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

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

  Possiblement sigui Pandora l'exemple més evident —però no pas l'únic— de la força de la cooperació. La National Library of Australia compta amb diverses institucions sòcies de projecte: l'Australian War Memorial, l'Australian Institute of Aboriginal and Torres Strait Islander Studies i les biblioteques dels diferents estats del país, entre d'altres. L'aliança proporciona rigor en la selecció, suport als pressupostos i presència mediàtica i en la comunitat de la recerca australiana.

Els principals inconvenients o punts febles del model selectiu són:

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

  El projecte britànic UKWA està liderat per la British Library i compta, entre altres socis importants, amb la National Library of Wales i amb la National Library of Scotland. El projecte està en desenvolupament, i els primers resultats han estat fets públics recentment (maig de 2005), però el fet és que alguns ítems seleccionats pels socis de projecte donen una visió parcial del web britànic. Per exemple, sota l'epígraf “teatre” l'únic ítem disponible és “theatre in Wales”.

- *Cost elevat*. La selecció, la gestió i el seguiment dels acords i les captures, i especialment l'anàlisi documental dels recursos, són tasques molt intensives que encareixen el cost per ítem en recursos humans. El fet que les institucions que gestionen els dipòsits siguin les biblioteques nacionals garanteix una alta qualitat en la descripció i indexació dels recursos, habitualment per llenguatge de metadades.

  En el congrés celebrat a Canberra el novembre de 2004,[9] la responsable del projecte australià Pandora desvetllava que el cost per a la gestió d'un ítem digital pot arribar a ser cinc vegades superior al d'una monografia.

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

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

2.1.3 El model híbrid

Els dos models anteriors han deixat pas a models híbrids que complementen la captura sistemàtica del web nacional (model típicament integral) amb acords amb institucions productores segons els interessos temàtics (model selectiu). Addicionalment, el projecte es pot dirigir a efectuar captures selectives de determinats esdeveniments d'interès general com, per exemple, els Jocs Olímpics, eleccions, catàstrofes naturals, etc.

El cas danès, per exemple, està enfocat en l'acció triple que suposa la captura exhaustiva del web danès, els acords amb entitats editores del país i la captura exhaustiva però focalitzada d'esdeveniments d'interès.

De l'estudi detallat dels dipòsits existents es desprèn que aquesta és la tendència cap a la qual tendeixen la majoria dels projectes integrals (Àustria, els Països Baixos, Suècia, Finlàndia, etc.).

Lògicament, els projectes híbrids incorporen alguns dels avantatges descrits anteriorment (col·lecció rica i equilibrada, màxim accés, impuls dels acords estratègics, compilació automatitzada i seguiment de les llacunes), però també elements negatius (cost elevat, equips més nombrosos o càrrega de gestió).

2.2 Dipòsits digitals nacionals

S'han trobat referències de vint casos de dipòsits nacionals:[10] Alemanya, Austràlia, Àustria, el Canadà, Dinamarca, els Estats Units d'Amèrica, Estònia, Finlàndia, França, Grècia, Islàndia, el Japó, Lituània, Noruega, Nova Zelanda, els Països Baixos, el Quebec, el Regne Unit, la República Txeca i Suècia. Atenent al model que segueixen, es poden classificar de la manera següent:

- Model integral (50 %): Alemanya, Àustria, Estònia, Finlàndia, Grècia, Islàndia, Lituània, Noruega, la República Txeca i Suècia.
- Model selectiu (35 %): Austràlia, el Canadà, els Estats Units d'Amèrica, el Japó, els Països Baixos, el Quebec i el Regne Unit.
- Model híbrid (15 %): els dipòsits de Dinamarca, França i Nova Zelanda es poden considerar projectes que s'escauen plenament dins del model híbrid.

Tanmateix, com ja s'ha esmentat, la major part dels dipòsits que segueixen un model integral han adoptat mesures per incloure determinats recursos (com ara publicacions periòdiques) que els fan acostar a paràmetres híbrids. Aquesta és la tendència generalitzada.

A continuació analitzarem els casos existents, centrant-nos en tres exemples que representen els models anunciats: integral (Kulturarw3 de Suècia), selectiu (Pandora d'Austràlia) i híbrid (Netarkivet de Dinamarca). Pel que fa a la resta de projectes, en un annex es fa una descripció sumària de les característiques de cadascun. No s'aporten dades d'Islàndia ni d'Estònia per manca de bibliografia.

2.2.1 Kulturarw3 (http://www.kb.se/kw3/ENG)

El projecte és liderat per la Kung. Royalbiblioteket (Suècia). S'inicia l'any 1996 i segueix el model integral. És exhaustiu pel que fa a la captura del web suec —350.000 webs

(febrer de 2005)— i la catalogació dels materials no és una prioritat. L'accés al fons està limitat a les dependències de la Kung. Royalbiblioteket.

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

La col·lecció cobreix revistes digitals i publicacions periòdiques no diàries, a excepció, des de fa uns mesos, d'una selecció de més de 100 títols de diaris suecs, documents estàtics (arxius electrònics) i documents dinàmics amb enllaços. Ulteriorment es recopila el contingut de llistes de discussió, i arxius FTP oberts. Per les dades fetes públiques el febrer de 2005, sabem que en aquella data les dimensions de Kulturarw3 eren de 306 milions d'arxius i uns 10.000 Gb: 350.000 llocs web.

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

Els punts forts del projecte Kulturarw3 són els derivats de l'exhaustivitat en la compilació automàtica i el valor pel fet de plasmar la societat digital sueca.[11] S'hi poden trobar des de revistes científiques a *weblogs* d'ONG. Els punts febles estan relacionats amb importants llacunes en el control d'allò que es captura, en la nul·la profunditat en la infranet (continguts de pagament o amb contrasenya, pàgines òrfenes, etc.) i en la manca de catalogació de l'arxiu. Com s'ha apuntat, el fet que l'arxiu només sigui consultable a les dependències de la biblioteca sueca és una característica negativa del sistema.

2.2.2 Pandora (http://Pandora.nla.gov.au/index.html)

El projecte és liderat per la National Library of Australia (Austràlia). S'inicia l'any 1996 i segueix el model selectiu. El seu abast se centra en la selecció de publicacions en línia i webs sobre Austràlia, d'autor australià o sobre un tema australià. La catalogació és exhaustiva, i les possibilitats de cerca, molt avançades. Disposa d'un programari propi, *Pandas*, que s'ha implementat en altres projectes.

L'arxiu web d'Austràlia, Pandora, va ser creat el 1996 per la National Library of Australia per garantir l'accés permanent a una selecció de publicacions en línia i de llocs web d'Austràlia i sobre Austràlia.

A falta d'una llei que reguli el dipòsit legal digital (la vigent és de 1968), la política de la biblioteca i els seus socis de projecte, que formen el comitè científic de la política selectiva, és arribar a acords amb les entitats editores dels documents susceptibles de ser capturats. S'ha publicat una guia dels criteris en què es basa la selecció dels llocs web capturats. Les dades estadístiques de setembre de 2005 mostren que l'arxiu conté 27 milions de fitxers i que té un creixement mensual de 30 Gb. És consultable en línia.

Els desavantatges del sistema australià estan relacionats amb la seva pròpia naturalesa:[12] el criteri de la selecció és forçosament subjectiu, malgrat la transparència de la política de selecció. El context (els enllaços als quals apunta el recurs), queden deslligats del document, perquè poden no estar inclosos en la selecció. Finalment, el cost de tractament (selecció, captura periòdica, catalogació, etc.) de cada ítem és molt elevat.

Per contra, els beneficis es concentren en la qualitat del tractament i la presentació del patrimoni. L'accessibilitat en línia, en règim obert, és possible pels acords subscrits amb els productors (que comporta l'accés als recursos de la infranet). Les dades de la catalogació són compartibles amb la resta d'equipaments australians (o internacionals). Es procura un creixement temàtic equilibrat de la col·lecció.

Vistos els models integral i selectiu, la tercera via que hem de considerar és la mixta. Com ja s'ha apuntat, bona part dels dipòsits digitals nacionals plantejats inicialment com a integrals han anat adoptant mesures per incloure recursos molt significatius, com ara publicacions periòdiques, en els seus fons.

Tres són els projectes pioners a apostar per una política clara de conjugació de les captures exhaustives, els acords amb institucions i organitzacions, i el detall per a activitats concretes: França, Dinamarca i Nova Zelanda. A continuació s'examina amb detall el projecte de Dinamarca.

2.2.3 Netarkivet (http://netarchive.dk/index-en.php)

El projecte és liderat per Det Kongelige Bibliotek (Dinamarca). S'inicia l'any 1998 i segueix el model híbrid que es basa en la captura exhaustiva, els acords per a la selecció i les activitats especials relacionades amb la realitat danesa. Les tasques de captura integral s'han iniciat el juliol de 2005. Des de 1997, la Llei de dipòsit legal inclou "totes" les publicacions de Dinamarca, i des de 2005 la biblioteca danesa té potestat per capturar tots els tipus de pàgines web daneses. La Kongelige Bibliotek facilita el lliurament del dipòsit legal per mitjà d'un formulari web.

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

El 24 de juny[13] de 2005 s'anuncia al gran públic la posada en marxa del projecte danès en el seu vessant integral, lligant-ho amb la nova modificació de la Llei de dipòsit legal.

Després de la primera fase de captura integral de juliol de 2005 amb el programari *Heritrix*, el projecte Netarkivet conté 600.000 dominis .dk, xifra que representa, segons la mateixa institució, el 60 % dels dominis danesos. Paral·lelament, fins a l'any 2004 i per al model selectiu, el dipòsit s'acosta als 500 terabytes de volum, amb un exponent de creixement anual de 30 Tb.

No és accessible en línia.

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

- Dinamarca és un país petit (5,4 milions d'habitants), amb un domini propi (.dk) i una llengua pròpia (danès); la selecció (i les gestions consegüents) és relativament senzilla perquè és limitada.
- Per la mateixa raó, també són aparentment senzilles les captures exhaustives. Per al desenvolupament del programari de captura selectiva i gestió, el Netarkivet ha col·laborat en el Nordic Web Archive (NWA), juntament amb la resta de biblioteques escandinaves.
- La legislació que afecta el dipòsit legal ha anat modificant-se (darrera revisió, al començament de juliol de 2005),[14] per ampliar-se a les publicacions digitals en línia. Un formulari en línia facilita la tasca dels productors i editors web.
- Hi ha dos socis destacats implicats en la coordinació del projecte: la biblioteca nacional i un centre de recerca d'una important universitat. S'elaboren acords estratègics sobre la base de determinades temàtiques.

2.3 Organitzacions i projectes suprainstitucionals

Els projectes descrits, que duen a terme les biblioteques nacionals, es troben sovint sota "paraigües" més amplis de cooperació entre biblioteques o altres tipus d'institucions.

2.3.1 International Internet Preservation Consortium

L'organització que aplega la major part d'aquestes iniciatives és l'IIPC (International Internet Preservation Consortium, http://netpreserve.org), que té la missió d'adquirir, preservar i fer accessibles el coneixement i la informació sobre Internet per a les futures generacions de tot el món, promovent l'intercanvi global i les relacions internacionals.

Va ser creat formalment el juliol de 2003 pels 12 membres que actualment formen el consorci: 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), Internet Archive (http://www.archive.org), 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).

Els objectius de l'IIPC són els següents:

- Permetre la recol·lecció d'una part rica del contingut d'Internet d'arreu del món, perquè sigui preservada de manera que pugui ser arxivada i preservada i assegurat l'accés en el temps.
- Fomentar el desenvolupament i l'ús d'eines comunes, tècniques i estàndards que permetin la creació d'arxius internacionals.
- Animar i donar suport a les biblioteques nacionals d'arreu per arxivar i preservar Internet.

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

El Consorci, doncs, no captura webs, sinó que agrupa una sèrie d'institucions que ho fan, i té com a objectiu promoure aquestes activitats.

2.3.2 Nordic Web Archive

L'NWA (Nordic Web Archive, http://nwa.nb.no) és un fòrum de les biblioteques nacionals escandinaves (Dinamarca, Finlàndia, Islàndia, Noruega i Suècia) per a la coordinació i l'intercanvi d'experiències en els camps de la captura i l'emmagatzematge de documents web.[15]

Des de novembre de 2000 s'ha desenvolupat el conjunt d'eines NWA:[16] un programari per accedir als document web arxivats, creat emprant PHP, Perl i Java, amb estàndards oberts com ara els protocols HTTP i XML per a la comunicació entre les diferents parts del sistema. L'ús del paquet de programari (cerca i navegació per l'arxiu web) es realitza per mitjà d'un cercador web estàndard, i no es necessari cap *plugin* específic.

La iniciativa va ser fundada per Nordunet2 (programa de recerca dels escandinaus), Nordinfo (consell escandinau per a la informació científica que inclou les biblioteques de recerca) i les biblioteques nacionals escandinaves.

2.3.3 Internet Archive

L'Internet Archive (http://www.archive.org) és una organització sense ànim de lucre fundada el 1996 per construir una “biblioteca d'Internet” i oferir accés permanent a

investigadors, historiadors, personal acadèmic i el públic en general a les col·leccions històriques en format digital. Situat a l'antiga presó de San Francisco (EUA), l'arxiu ha rebut donacions d'IBM, d'Alexa (filial d'Amazon) i d'altres organitzacions que han facilitat el seu creixement.

Rep el suport de diversos organismes, com ara The Library of Congress, els US National Archives i els UK National Archives, entre d'altres.

A l'actualitat, Internet Archive es considera l'arxiu web més gran del món,[17] i inclou text, àudio, imatge en moviment i programari, així com pàgines web arxivades de tot el món, amb un nombre representatiu de recursos catalans.[18] En accés obert i en línia, el gegant conté en un petabyte un total aproximat de 600 milions de llocs web, des de 1996 fins a l'actualitat, i cada dos mesos es realitza una captura massiva que afecta milions de pàgines web (creixement mensual de 20 Tbytes), seguint el model exhaustiu que Suècia i altres països representen.

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

Recentment i amb seu a Amsterdam s'ha creat l'European Digital Archive, branca europea de l'Internet Archive.

3 El futur és generalitzat, híbrid, costós i cooperatiu: conclusions sobre la panoràmica

L'interès per la preservació digital està ja generalitzat als països desenvolupats, encara que amb un grau de desenvolupament heterogeni. Probablement, quan es publiqui el present article existiran més dels vint projectes mencionats, encara que estiguin en fase de disseny.

El futur és híbrid: la diferenciació en models (integral *versus* selectiu) representa només la primera fase de desenvolupament dels projectes.

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

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

Els acords amb els productors i editors web són garantia d'èxit. No sempre una llei moderna de dipòsit legal acompanya els dipòsits digitals nacionals que existeixen, i l'accés a les infranets (i a la Internet invisible) ha de preveure's, amb llei o sense.

A Catalunya, la Biblioteca de Catalunya ha iniciat amb el seu projecte PADICAT (Patrimoni Digital de Catalunya) les actuacions necessàries per fer sempre accessibles les webs catalanes.

Data de recepció: 03/10/2005. Data d'acceptació: 25/10/2005.

Annex. Descripció sumària dels projectes de dipòsits nacionals

AOLA (http://www.ifs.tuwien.ac.at/~aola)

- Liderat per l'Österreichische National-bibliothek (Àustria).
- Iniciat el 1999.
- Model integral amb elements del model híbrid.
- El creixement previst és de 7 Gb diaris.
- El projecte ha patit aturades per manca de fons.
- El programari NEDLIB de la primera fase va donar lloc al COMBINE en una etapa posterior.

[*Archive of Czech web resources*] (http://webarchiv.nkp.cz/index-e.html)

- Liderat per la Národní knihovna Ceské Republiky (República Txeca).
- Iniciat el 2000.
- Model integral.
- En col·laboració amb altres institucions bibliotecàries i de recerca, la biblioteca nacional txeca ha impulsat les captures anuals del domini .cz per mitjà d'una adaptació del programari NEDLIB.
- S'ha previst incloure publicacions digitals en etapes futures.

[*Archiving the French web*] (http://www.bnf.fr)

- Liderat per la Bibliothèque Nationale de France (França).
- Iniciat el 2000.
- Model híbrid.
- L'abast del projecte inclou captures automàtiques a gran escala, captures sistemàtiques i contínues d'una selecció de llocs web (el 10 % del total), dipòsit de la infranet i captures temàtiques de llocs web molt efímers (eleccions franceses de 2002: 1.900 llocs web).

[*Archiving the Web Greek*]

- Liderat per l'Athens University of Economics and Business (Grècia).
- Iniciat el 2003.
- Model integral.
- El projecte grec és un experiment destinat a capturar el domini grec, amb programari propi.

Deposit.ddb.de (http://deposit.ddb.de/online/vdr/titel.htm)

- 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 (http://epe.lac-bac.gc.ca/)

- Liderat per Libraries and Archives Canada (el Canadà).
- Iniciat el 1994.
- Model selectiu.
- A partir de l'EPPP (*Electronic Publications Pilot Project*), de 1994–95, es va crear l'E-Collection, destinat a l'arxiu en línia de publicacions digitals, a text complet.
- L'actualització del projecte, 2004–05, inclou tesis en línia, webs, etc.

e-Depot (http://www.kb.nl/dnp/e-depot/e-depot-en.html)

- Liderat per la Koninklijike Bibliotheek (els Països Baixos).
- Iniciat el 1995.
- Model selectiu.
- A partir dels acords amb els editors (als Països Baixos s'ubiquen un nombre important de multinacionals editores), s'apunta a les revistes publicades a Holanda, on no existeix legislació estricta de dipòsit legal.
- Conté tres milions de números de revistes (març de 2005).

EVA (http://www.lib.helsinki.fi/tietolinja/0203/webarchive.html)

- Liderat per la Helsingin yliopiston kirjasto (Finlàndia).
- Iniciat el 1997.
- Model integral amb elements del model híbrid.
- La biblioteca nacional de Finlàndia lidera el NWA (Nordic Web Archive), que pretén ser l'arxiu web escandinau.
- El projecte EVA, adreçat en successives etapes a publicacions periòdiques, va donar pas el 2001 a l'arxiu web que inclou el domini .fi.

IRIS (http://catalogue.bnquebec.ca:4400/cap_fr.html)

- Liderat per la Bibliothèque nationale du Québec (el Quebec).
- Iniciat el 2000.
- Model selectiu.
- 3.469 monografies i 1.100 títols de revistes digitals (octubre de 2004) formen el cos de l'arxiu, que està integrat en el catàleg IRIS.
- Captura mitjançant acords amb el Govern quebequès, en una primera fase.

LIBIS Electronic Resources Subsystem (http://www.libis.lt/en/welcome.html)

- Liderat per la Martynas Mazvydas (Lituània).
- Iniciat el 2002.
- Model integral.
- El projecte LIBIS consisteix a completar el catàleg LIBIS amb les captures procedents del sistema NEDLIB.
- El projecte inclou metadades Dublin Core.

Minerva (http://www.loc.gov/minerva/)

- Liderat per The Library of Congress (els Estats Units d'Amèrica).
- Iniciat el 2000.

- Model selectiu temàtic.
- Associat al gegant Internet Archive, el recurs Minerva captura selectivament 35 llocs web.
- En les eleccions presidencials de 2000 (i en altres dates especials, com ara l'11-S) s'augmenta la captura selectiva.
- S'ha previst incloure-hi publicacions digitals en etapes futures.

New Zealand's digital Heritage (http://www.natlib.govt.nz/bin/media/pr?item=1085885702)

- Liderat per la National Library of New Zealand (Nova Zelanda).
- Iniciat el 1999.
- Model híbrid.
- La llei de dipòsit legal novazelandesa, de 2003, obria el panorama als recursos en línia, incloses les publicacions en règim obert i també la infranet.
- Pressupost global del patrimoni digital (2004): 14 M€.
- Empra el programari Pandas, però l'accés no és obert.

Paradigma (http://www.nb.no/paradigma/eng_index.html)

- Liderat per la Nasjonalbiblioteket (Noruega).
- Iniciat el 2001.
- Model integral amb elements del model híbrid.
- Amb captures anuals, a partir de 2003 s'hi va incloure la captura dels dominis internacionals amb contingut noruec, així com 65 diaris digitals.
- La indexació dels recursos es produeix automàticament per mitjà del programari FAST.

UK Web Archive (http://www.webarchive.org.uk/)

- Liderat per la British Library (el Regne Unit).
- Iniciat el 2004.
- Model selectiu.
- Seguint el model australià, amb el mateix programari Pandas i una interfície de consulta molt similar, el maig de 2005 presentava 1.030 llocs web amb els quals s'ha arribat a acords.
- En la selecció, i en tot el projecte, participen la resta de biblioteques nacionals del Regne Unit.

WARP (http://warp.ndl.go.jp)

- Liderat per la National Diet Library (el Japó).
- Iniciat el 2002.
- Model selectiu.
- Les esmenes a la llei de dipòsit legal de 2000 inclouen CDR i altres materials digitals en suports físics.
- La biblioteca nacional del Japó recull (juny de 2004), amb acords, 600 llocs web (Administració, universitats, empreses, etc.) i 110 diaris electrònics.

Notes

[1] *Directrices para la preservación del patrimonio digital* (Canberra: Unesco, 2003), <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 03/10/2005].

[2] L'UK Web Archiving Consortium fixa en 44 dies la mitjana de vida d'una pàgina web <http://info.webarchive.org.uk/pressrelease21-06-04.html>. [Consulta: 03/10/2005].

[3] El 1999 la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona va tenir una iniciativa en aquesta mateixa línia en organitzar un seminari sobre aquesta qüestió i publicar-ne els treballs: *Biblioteques digitals i dipòsits nacionals de recursos digitals* (Barcelona: Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, 1999).

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

[5] José Antonio Cordón, "El depósito legal y los recursos digitales en línea". En: *Las bibliotecas nacionales del siglo XXI* (Valencia: Biblioteca Valenciana, 2005), <http://bv.gva.es/documentos/Ponencias/Cordon.pdf>. [Consulta: 03/10/2005].

[6] La traducció de *nu* és *ara*. Malgrat que sigui el domini geogràfic de l'illa Nieu, a la Polinèsia, en suec i en altres llengües escandinaves és molt utilitzat per dotar d'un component dinàmic el nom del domini.

[7] Eines com ara Maxmind (http://www.maxmind.com) o Ip2location (http://www.ip2location.com) faciliten la ubicació geogràfica dels servidors.

[8] InternetLab (http://internetlab.cindoc.csic.es) pertany al CINDOC-CSIC.

[9] *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[10] Un recurs que proporciona informació detallada dels projectes nacionals és *PADI: Preserving Access to Digital Information* (http://www.nla.gov.au/padi/) de la National Library of Australia. PADI conté informació actualitzada de la pràctica totalitat dels projectes existents, així com informació diversa dels aspectes relacionats amb la preservació web (dipòsit legal, eines, bibliografia, etc.). La National Library of Australia també va organitzar, el novembre de 2004, el congrés Archiving web resources, en el qual són accessibles en règim obert la major part de presentacions realitzades en aquella activitat. Finalment, també cal tenir present l'IWAW (International Web Archiving Workshop, http://www.iwaw.net) que se celebra anualment, organitzat per un grup de professionals procedents dels diversos projectes. En aquest espai web també es pot consultar la documentació relativa als projectes existents. El creixement del nombre de projectes destinats a crear dipòsits nacionals és constant, i pròximament s'hauran de sumar als exposats en aquest article els de les biblioteques nacionals de Luxemburg, Singapur, Egipte, Croàcia i, evidentment, el de la Biblioteca de Catalunya.

[11] Joan Mannerheim, "Collect all, catalogue some". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[12] Margaret Phillips, "What to collect and how to do it: the National Library of Australia's selective approach". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[13] La data té una càrrega simbòlica: la festa de Sant Joan (el solstici d'estiu) és especialment celebrada als països escandinaus.

[14] "New legal deposit law", *Netarchive.dk*, <http://netarchive.dk/newsite/news/index-en.php>. [Consulta: 03/10/2005].

[15] Porsteinn Hallgrímsson, Sverre Bang, "Nordic Web Archive". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[16] El paquet de programari es pot descarregar a Sourceforge (http://nwatoolset.sourceforge.net).

[17] Michele Kimpton, "Saving the web for future generations". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[18] En són exemples (agost de 2005) les pàgines web de la Generalitat de Catalunya (http://www.gencat.net), amb 207 captures des de maig de 2002; el diari *Avui* (http://www.avui.es), amb 55 captures des de gener de 1998, o la Universitat de Barcelona (http://www.ub.es i http://www.ub.edu), amb 223 captures des de febrer de 1997, etc.