bid textos universitaris de biblioteconomia i documentació
número 16
juny de 2006
ISSN 1575-5886
DL B-19.675-1998
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Programari de codi lliure per gestionar dipòsits digitals: el procés de tria dut a terme al CBUC

[Traducción automática al español]

ANTONI BORRÀS
Universitat Pompeu Fabra
antoni.borras@upf.edu

JUAN CARREÑO
Universitat de Girona
jcarreno@pas.udg.es

FERRAN JORBA
Universitat Autònoma de Barcelona
Ferran.Jorba@uab.es

JORDI PRATS
Universitat Politècnica de Catalunya
jordi.prats@upc.es

RAMON ROS
Consorci de Biblioteques Universitàries de Catalunya
rros@cbuc.es

Resum [Abstract] [Resumen]

Es descriu el procés de selecció dut a terme en el Consorci de Biblioteques Universitàries de Catalunya (CBUC) durant el 2004 per tal de seleccionar el programa per gestionar els dipòsits digitals del CBUC. Dues de les condicions per a la selecció eren que es tractés de programari lliure i, a més, que es pogués implementar de manera immediata. Els programes analitzats van ser *ARNO*, *CDSware*, *DSpace*, *EPrints*, *Fedora*, *i-Tor* i *MyCoRe*. En el text es descriuen les principals característiques de cada un d'aquests programes i es justifica l'elecció de *DSpace*.

1 Introducció

Entre els objectius del Consorci de Biblioteques Universitàries de Catalunya (CBUC) per al 2004 hi havia el de "posar en funcionament un dipòsit de documents digitals". Com a fase prèvia, calia seleccionar un programari que el pogués gestionar. El procediment de treball per triar-lo va ser crear un grup d'informàtics i bibliotecaris que analitzessin i provessin les diferents opcions existents en aquell moment amb diferents condicionants; entre aquests, que fos de codi lliure, que fos compatible amb l'estàndard OAI-PMH[1] i que es pogués posar en marxa d'una manera ràpida, sense la necessitat d'esperar possibles desenvolupaments locals. Aquest article, escrit per gairebé tots els membres d'aquell grup, resumeix el treball realitzat durant les diferents reunions i proves i raona la selecció de *DSpace* com el programari per gestionar els dipòsits digitals del CBUC.

1.1 L'encàrrec del CBUC

Davant l'important creixement que hi ha hagut els darrers anys en la quantitat d'informació electrònica que han de gestionar les biblioteques, sorgeix la necessitat de tenir un sistema que hi doni suport i en garanteixi l'accés i la preservació futura. Així, l'any 2003 les institucions integrades en el Consorci de Biblioteques de Catalunya (CBUC) veuen la necessitat de crear un dipòsit comú per a aquest tipus d'informació, i això queda reflectit en el pla de treball del 2004, que fa la proposta de "posar en funcionament un dipòsit de documents digitals" i, per gestionar-lo, usar un mateix sistema informàtic.

Aquest dipòsit hauria de permetre que els membres del CBUC poguessin aconseguir els tres objectius bàsics següents:

- Experimentar els processos de captació i preservació d'objectes digitals.
- Concretar alguns paràmetres que facilitin que els diferents objectes digitals siguin consultables ara i en el futur de manera conjunta (per exemple, metadades).
- Crear una comunitat d'usuaris que comparteixi la mateixa solució informàtica per al mateix problema.

El sistema triat havia de permetre que els documents emmagatzemats en aquests dipòsits complissin els requisits següents:

- Que siguin accessibles directament, a més de poder ser referenciats des de programes o aplicacions diversos (a través d'una adreça estable).
- Que puguin ser sotmesos a processos que permetin, si escau, accessos restringits o filtrats.
- Que puguin ser sotmesos a processos que n'assegurin o en facilitin la usabilitat futura (preservació).

Al mateix temps que hi havia una idea més o menys concreta del que havia de ser aquest dipòsit, també es veia clar el que no havia de ser: ni un sistema d'aprenentatge virtual, ni un catàleg automatitzat de biblioteca, ni tampoc un sistema de gestió de la recerca feta a cada universitat. Això permetia concentrar-se més en el producte demanat i no interferir en altres projectes ja en curs que tenien les seves pròpies solucions informàtiques.

Tot i que en aquell moment ja existien sistemes de gestió de dipòsits comercials, la necessitat de moltes biblioteques de trobar una solució havia fet que existís un nombre important de productes de codi lliure d'una qualitat equiparable a la dels comercials. Al mateix temps, la necessitat immediata de trobar i utilitzar el producte, la voluntat del CBUC d'utilitzar programari lliure sempre que fos possible i el marc temporal en què es volia prendre una decisió, van fer que l'encàrrec especifiqués que es volia una solució d'aquest tipus.

Així, doncs, la Comissió Tècnica del CBUC va crear un grup de treball format per informàtics i bibliotecaris de diferents institucions perquè busquessin, analitzessin i, finalment, fessin una proposta de programari per gestionar un dipòsit d'aquestes condicions.

1.2 Procediment de treball

El grup creat havia de fer una proposta en un termini de dos mesos, ja que es volia crear aquest dipòsit de manera immediata; per tant, entre altres condicionants, el programari havia d'estar completament acabat (no podien ser versions beta), havia de ser utilitzable directament per l'usuari final (no podia ser una eina de gestió interna de les biblioteques) i havia d'haver estat provat i en funcionament en altres biblioteques del món.

Com a punt de partida, es va decidir centrar-se en la primera versió d'un estudi de l'Open Society Institute (OSI) sobre programari de codi lliure per gestionar dipòsits institucionals.[2] Durant el curs d'aquests estudis va aparèixer una segona versió que també es va incorporar. Aquest document recollia l'estat de l'art d'aquests gestors amb les

restriccions principals d'haver de ser de codi lliure i haver de complir les especificacions del protocol per a l'intercanvi de metadades OAI-PMH, versió 2. Aquestes condicions encaixaven plenament en l'encàrrec que s'havia fet al grup.

L'informe de l'OSI avalua set sistemes diferents: *ARNO*, *DSpace*, *EPrints*, *Fedora*, *CDSware*, *i-Tor* i *MyCoRe*, i exposa les diferents característiques de cadascun. Alguns sistemes coneguts, com ara *Greenstone*, no es van avaluar, ja que no apareixien a l'informe; el motiu és que en aquell moment encara no complien l'estàndard OAI-PMH. El grup de treball, després de diferents reunions i activitat conjunta per mitjà d'un wiki, va decidir centrar-se en els deu punts que es consideraven més importants en el procés de tria amb l'objectiu demanat. Eren els següents:

- Restricció d'accés per tipus d'usuari.
- Autenticació (LDAP, usuari i contrasenya, adreces IP, etc.).
- Restricció d'accés quant a l'objecte i/o la col·lecció.
- Possibilitat de definir col·leccions múltiples dins la mateixa instància del sistema.
- Existència d'una pàgina principal per a cada col·lecció.
- Sistema configurable dels papers de lliurament dels objectes digitals.
- Estadístiques en general.
- Càrregues en massa, eines convertidores, etc.
- Esquema bàsic de metadades.
- Funcionalitat de cerca, interfície d'usuari i els tipus de cerques que pot fer.
- Suport a la preservació.

Es va fer un estudi sobre paper dels diferents sistemes i es van descartar, en una primera fase, els sistemes *ARNO*, *i-Tor* i *MyCoRe*. Després se'n van instal·lar alguns, es van estudiar aquests deu punts i es va descartar el sistema *Fedora*. Finalment, es va passar a la fase d'estudi final i en profunditat dels sistemes *CDSware*, *EPrints* i *DSpace*.

2 Sistemes estudiats

2.1 ARNO

ARNO (*Academic Research in the Netherlands Online*), <http://www.uba.uva.nl/arno>. El projecte *ARNO* va ser fundat per IWI (Innovation in Scientific Information Supply) i hi han participat les universitats d'Amsterdam, Tilburg i Twente, i està pensat per dipositar-hi i fer disponible la producció científica de les institucions participants. En el moment en què aquest producte va ser avaluat, el gestor de bases de dades havia de ser necessàriament *Oracle*, mentre que les nostres preferències eren utilitzar altres gestors de bases de dades de codi obert com poden ser *MySQL* o *PostgreSQL*. D'altra banda, *ARNO* per si sol no incorpora cap interfície d'usuari, sinó que caldria implementar-la amb algun programari d'un tercer. Tampoc no incorpora cap estratègia de preservació de la informació digital. Finalment, el fet que el nombre d'instal·lacions fos molt baix i que estigués restringit a Holanda va fer que desestiméssim aquesta aplicació ja en la primera fase.

Figura 1. Pàgina de resultats d'una cerca feta amb *ARNO*

2.2 CDSware

CDSware, <http://cdsware.cern.ch/>, és el projecte d'unificació de diferents aplicacions del *CERN*[3] entorn de la gestió de la seva col·lecció digital de documents, escrits al llarg del temps per diferents equips de persones, amb diversos llenguatges de programació i estils. Destaca inicialment pel seu suport nadiu al format MARC21, la seva interfície multilingüe i multialfabet (en la versió més recent, 0.7.1, 11 idiomes i 3 alfabets), i unes possibilitats de cerca molt riques (per índex o per paraula clau, expressions regulars, cerques aproximades, registres semblants, etc.), amb uns temps de resposta rapidíssims, malgrat les dimensions de la base de dades, com ara la del mateix CERN, <http://cdsweb.cern.ch/>, de més de 800.000 registres bibliogràfics.

Durant el període de proves vam avaluar les versions *de desenvolupament* 0.3.0 i 0.3.1, tal com recomanava la documentació per a noves instal·lacions. I va resultar que era un paquet, efectivament, encara poc recomanable per a institucions que volguessin una aplicació acabada. Pel que fa als requeriments informàtics, tot i que a la llarga serà una aplicació basada en Apache, MySQL i Python, encara hi havia, en la versió 0.3.x, mòduls importants amb PHP que, en versions posteriors, han anat retirant a favor de Python. Això feia que, d'una banda, la llista de prerequisits d'instal·lació fos força gran (Apache2, MySQL, Python i mod_python, PHP, Perl, WML, xpdf, wvWare, xlhtml, etc.),[4] i de l'altra, es veia que encara estaven treballant en una millor integració i unificació d'aquests diferents mòduls.

Vam fer les instal·lacions de prova en la distribució de Linux anomenada *Debian*, i les dificultats més grans van consistir, en primer lloc, en els requisits de PHP, que no eren els que venien per defecte, i a adaptar-nos a un entorn de treball Unicode, concretament en UTF-8,[5] així com la complexitat i les exigències de configuració del servidor web.

Un cop instal·lada l'aplicació, vam crear diferents usuaris amb permisos d'administració, consulta, etc., i hi vam afegir col·leccions reals i virtuals. Finalment, vam fer proves de lliurament i aprovació de documents, a més

d'intentar, sense èxit, modificar o crear formularis.

Els registres bibliogràfics poden ser de molt diversa tipologia (per exemple, al *CERN* hi ha *preprints*, conferències, congressos, col·leccions històriques de cartes, fotografies o vídeos); *CDSware* els agrupa jeràrquicament en col·leccions i subcol·leccions, reals o virtuals, i sense limitació de nivells. Una col·lecció és un conjunt de registres que compleix una condició determinada: en general, les col·leccions reals es defineixen per un valor del subcamp 980$a;[6] en aquest cas, un registre que pertanyi a una col·lecció no pertany a cap altra. Ara bé, es poden definir paral·lelament col·leccions *virtuals*, o de conjunts de registres que tinguin una altra característica transversal (p. ex., un autor, una llengua o una temàtica).

Figura 2. Pàgina de resultats d'una cerca feta amb *CDSware*

El fet que *CDSware* tingui el MARC21 com a format bibliogràfic intern (era l'única de les aplicacions avaluades que el tenia) obre, evidentment, unes possibilitats de gestió de la informació bibliogràfica que van més enllà de Dublin Core, però també augmenta la complexitat de la seva configuració, sobretot perquè contínuament hi ha transformacions de format:

- D'una banda, el *CERN* no utilitza el MARC21 pur, sinó una adaptació,[7] i fer que *CDSware* s'avingui a l'estàndard requereix no solament una feina de canvi de paràmetres, sinó també, en algun cas, retocar alguna línia del codi del programa. Per tant, també cal modificar bona part dels paràmetres preconfigurats per defecte.

- A l'hora de definir els formularis de catalogació (especialment si els omplen els mateixos autors),[8] també cal configurar uns fitxers de conversió entre els noms dels camps i els valors MARC21 equivalents, una conversió que sempre serà aproximada.

- A la inversa, la majoria dels clients OAI-PMH només entenen Dublin Core, o sigui, que hi ha una altra conversió de sortida —en aquest cas amb pèrdua— a aquest format si es vol utilitzar el servidor OAI de *CDSware*.

- A partir dels camps i subcamps de MARC21 cal configurar els formats de sortida o presentació dels registres en format breu i complet, opcionalment enriquits amb enllaços, tant interns (a altres registres de la base de dades) com externs (a URL de fora del sistema).

Però on els membres de l'equip d'avaluació ens vam trobar amb les dificultats més serioses va ser en la gestió dels formularis de catalogació i en el circuit de treball. Va ser aleshores que ens vam veure incapaços d'entendre'n el funcionament per crear un circuit de catalogació complet, perquè segurament aquest és el mòdul menys obvi i amb la documentació menys clara. En resum, el sistema era d'una complexitat excessiva i no prou madur, de manera que no complia els requisits del CBUC.

Com en el cas d'altres aplicacions que s'han escrit en primer lloc per a usos interns en una institució i després s'han publicat amb llicència lliure, hi ha un temps en què l'aplicació va deixant de ser particular i es va generalitzant; però això, a part dels mèrits inicials, depèn també en bona part de la comunitat d'usuaris, desenvolupadors i contribuïdors externs, que hi van afegint necessitats diferents, suggerint solucions o aportant millores. Sense una voluntat i un esforç d'abstracció mutus, l'aplicació difícilment seria d'ús general. I, tot i que l'equip de desenvolupament del *CERN* tingui mires àmplies, *CDSware* està encara —però sobretot ho estava en el moment de fer-ne l'avaluació— en aquest període de generalització. Les necessitats del *CERN* són tan específiques, i els usuaris externs són encara pocs, que les instal·lacions fora del seu entorn originari requereixen una voluntat d'implicar-s'hi que no tothom pot assumir. El seu número de versió, encara inferior a l'1.0, és un reconeixement que està en aquest període de consolidació.

A partir de la versió 0.7.1, *CDSware* canviarà el seu nom a *CDS Invenio*.

2.3 DSpace

El projecte *DSpace*, <http://www.dspace.org/>, va ser desenvolupat conjuntament per les biblioteques del MIT (Massachusetts Institute of Technology, (http://libraries.mit.edu/) i l'empresa Hewlett-Packard (http://www.hp.com/). El seu objectiu és satisfer les diferents necessitats de difusió, organització i preservació dels objectes digitals: tant de dipòsits institucionals com de dipòsits d'objectes d'aprenentatge, o bé per a la gestió de recursos digitals. Actualment, *DSpace* és la segona aplicació més estesa (després d'*EPrints*), amb una comunitat d'usuaris molt gran[9] i amb institucions importants, entre elles unes quantes universitats de prestigi internacional, que la utilitzen, agrupades informalment en el que s'anomena la *DSpace Federation* (http://dspace.org/federation/).

Per efectuar les proves, vam utilitzar una instal·lació que ja estava funcionant a la Universitat Oberta de Catalunya (UOC). Es tractava de la versió 1.1 (actualment està disponible la versió 1.4). Pel que fa a la instal·lació, tot i que no es va partir des de zero, vam comprovar com el procés informàtic és relativament senzill: només fan falta els productes Apache, Java i Tomcat, a més de la base de dades *PostgreSQL* (l'únic gestor de bases de dades que estava suportat amb fiabilitat en aquell moment) sobre un entorn *Linux*.

Per catalogar els objectes digitals, *DSpace* disposa d'un conjunt de metadades en format Dublin Core, que vam considerar que seria suficient per assolir els nostres objectius.

Quant a aspectes favorables, cal destacar, per sobre de tot, la facilitat d'ús, tant per a l'administrador com per a l'usuari que incorpora documents, o per a l'usuari final que els consulta.

L'administrador pot gestionar completament tots els aspectes de l'aplicació des de la seva pròpia interfície web: creació de comunitats, col·leccions, assignació de permisos i grups, creació d'usuaris, regles, controls dels fluxos de treball, o bé delegar responsabilitats en determinats usuaris si ho considera convenient; com a exemples, la modificació dels camps Dublin Core, la desviació de fluxos, el control de formats per dipositar, etc. És a dir, en comptades ocasions l'administrador del sistema ha de baixar al nivell de sistema operatiu per realitzar processos, llevat de la càrrega massiva o importació d'una gran quantitat d'objectes, sempre que els fitxers de configuració de

l'aplicació estiguin ajustats racionalment a la mida del dipòsit i al tipus de material que s'ha de preservar. Totes aquestes característiques fan que el manteniment pugui ser dut a terme per personal no necessàriament informàtic.

En segon lloc, la incorporació de documents al dipòsit la pot realitzar qualsevol usuari autoritzat a través d'uns formularis molt senzills amb un sistema d'ajuts guiats pas a pas, semblant als mètodes utilitzats per comprar en els llocs web comercials. Bàsicament cal incloure els descriptors, el títol, el resum i, finalment, adjuntar el document. Quan l'usuari ha carregat un document al dipòsit, el sistema notifica automàticament la seva incorporació a l'editor de la col·lecció. Aleshores, aquest revisa o delega en tercers la validació o no del document i, un cop finalitzat el procés de revisió, i en cas que sigui acceptat, el sistema l'incorpora de manera definitiva a la base de dades. Finalment, *DSpace* notifica automàticament a l'autor l'acceptació o no del document i, en cas afirmatiu, informa els subscriptors de la col·lecció de la novetat dins la base de dades.

En darrer lloc, a l'usuari final li és molt fàcil i intuïtiu moure's per la interfície amb unes cerques per paraula clau, autor i títol prou potents. L'usuari disposa, ja des del moment en què es dóna d'alta, d'una sèrie de serveis addicionals com poden ser les alertes per correu electrònic o la subscripció a les novetats.

El problema més greu amb el qual ens vam trobar va ser el tractament dels diacrítics: en alguns productes de gestió documental no es tenen en compte les equivalències a l'hora de fer la cerca dels caràcters que tenen diacrítics (accents, per exemple) respecte dels que no en tenen, i això fa que l'usuari hagi d'escriure la seva consulta exactament tal com està introduït el text a la base de dades. En alguns casos, sobretot en llengües estrangeres, això és difícil i és probable que s'estiguin perdent resultats. Cal que el programa ignori els diacrítics tant quan busca com quan retorna els resultats.

Figura 3. Pàgina de resultats d'una cerca en el dipòsit RECERCAT, elaborat amb *DSpace*

Tant en la versió provada com en l'ultima versió instal·lada, les poques possibilitats de "jugar" amb les llengües són un problema rellevant: o bé s'opta per la traducció dels paràgrafs en anglès, o bé s'han d'establir al nivell local taules de conversió de paraules d'un idioma a un altre per preveure diferents idiomes alhora. És una feina considerable que s'havia de fer si es volia tenir un dipòsit basat en *DSpace* i en una llengua no anglesa (l'originària). Segons els desenvolupadors del producte, la propera versió ja inclourà característiques per fer una traducció més fàcil (fitxers o taules de llengua externs).

Finalment, un altre dels aspectes que cal millorar és el gràfic: un web estructurat en tres marcs, un de superior, un a l'esquerra i un de central és un model ja massa vist a Internet. És per això que, per tal de donar-hi un aspecte més modern, s'han de treballar una mica els estils i les visualitzacions de tot el conjunt. En tot cas, però, aquest és un element menor entre altres aspectes que ja hem considerat, altament positius.

Dels sis sistemes estudiats, *DSpace* no era pas el més avançat tècnicament, tampoc no era el que oferia més potencialitat per a l'usuari o per a l'administrador, com tampoc no era el més ràpid. Ara bé, en tots aquests aspectes sí que tenia una nota alta: era un sistema molt acabat, amb prestacions que superaven els mínims requerits, amb una comunitat alta d'usuaris. De tots els sistemes, era el que s'ajustava millor als requeriments demanats al grup i per això va ser proposat com a finalista.

2.4 EPrints

El projecte *EPrints*, <http://www.eprints.org>, neix a la University of Southampton l'any 2000 de la mà d'un projecte dirigit pel professor Stephen Harnad. Desenvolupat inicialment per Rob Tansley i després per Christopher Gutteridge, va trobar la primera implementació en el projecte *Crogprints*, un dipòsit obert temàtic especialitzat en psicologia, neurociència i lingüística. Actualment és l'aplicació per a la gestió de dipòsits oberts amb més instal·lacions, tant pel què fa a dipòsits institucionals com pel que fa als temàtics. Si també tenim en compte que la University of Southampton ha destacat pel fet de ser un dels nuclis de desenvolupament tecnològic més importants a l'entorn dels arxius oberts, amb projectes com ara *CiteBase*[10] o el *Registry of Open Access Repositories,*[11] no ha d'estranyar que, d'entrada, es presentés com un dels candidats amb més punts per ser seleccionat.

El producte està desenvolupat en llenguatges Perl i MySQL i es presenta com una aplicació d'instal·lació fàcil. Cal destacar que no són necessaris uns coneixements informàtics molt especialitzats per realitzar la instal·lació de l'aplicació, i la seva posada en funcionament es pot considerar quasi immediata. Una altra qüestió són els canvis que es vulguin fer en la parametrització inicial, o els aspectes vinculats a la personalització de la interfície d'usuari. En aquests casos, *EPrints* requereix sovint editar directament els arxius de configuració del sistema, i no proporciona una interfície amigable per realitzar aquests processos.

En el procés d'anàlisi dels productes es va poder constatar que *EPrints* complia molts dels requisits que s'havien prioritzat. Aspectes com ara les opcions de cerca, la flexibilitat en el disseny de l'estructura de metadades (Dublin Core) o un mínim suport en la gestió de la integritat de les dades estan previstos en la instal·lació bàsica del producte. La seva estructura de col·lecció única proporciona una interfície d'usuari simple i intuïtiva, a la qual cal afegir el suport multilingüe, així com la possibilitat de personalitzar-la si es tenen els coneixements adequats.

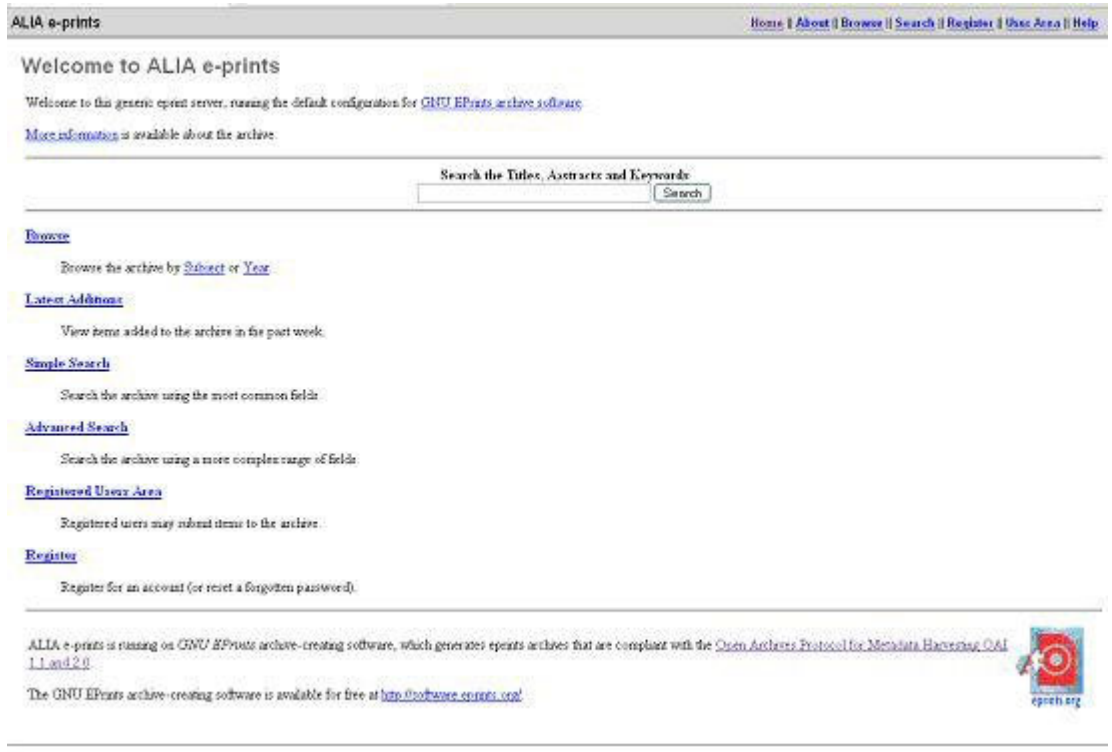

Figura 4. Pàgina principal d'un dipòsit desenvolupat amb *EPrints*

Cal destacar, també, altres aspectes no considerats com a punts prioritaris, però certament rellevants, com ara la possibilitat d'indexació dels objectes sobre la base de llenguatges controlats, o la flexibilitat a l'hora de carregar i gestionar en el sistema objectes complexos, com poden ser petits webs. Com a prova, es va fer la càrrega i indexació del web principal d'una de les universitats participants en el grup de treball, amb resultats molt satisfactoris.

No obstant això, aspectes importants en el moment d'iniciar un projecte consorciat, com ara la possibilitat de segmentar els continguts del dipòsit en diferents col·leccions o la poca flexibilitat en la configuració de rols i circuits de treball, van fer que s'anés desestimant aquesta opció. Hi podem afegir, com ja s'ha dit, que en la versió analitzada la interfície web d'administració no donava gaires opcions, i el manteniment del sistema depenia molt de la intervenció de personal informàtic.

Malgrat no haver estat seleccionada, *EPrints* és una aplicació prou compacta i acabada per poder oferir servei amb un cost de temps i desenvolupament prou raonable. Les característiques del projecte pel qual es va iniciar el procés de selecció van desaconsellar la tria d'aquesta aplicació, fet que no vol dir que no es pugui tenir en consideració en altres models de dipòsits o en projectes d'altres característiques. Les dificultats que s'han comentat quant a modificar els paràmetres de configuració del sistema queden minimitzades amb una parametrització inicial molt raonable.

2.5 Fedora

El projecte *Fedora*, <http://www.fedora.info/>, està desenvolupat per dues universitats americanes, la Cornell University Information Science i la University of Virginia Library, amb el suport econòmic de l'Andrew W. Mellon Foundation. Per valorar la solidesa del projecte, ens hem de situar necessàriament en el moment de l'estudi, al començament de 2004. Des de llavors, *Fedora* ha evolucionat i s'ha consolidat. En el moment d'iniciar els estudis, vam tenir l'ocasió de provar la versió 1.2, que només donava suport a sistemes *Linux*. En aquests

moments (abril de 2006), la versió vigent és la 2.1, sobre sistemes *Unix/Linux* i *Windows*, amb versions estables en ambdues plataformes. Encara avui continua tenint, entre els programaris de gestió de continguts digitals, trets distintius, com ho comentarem tot seguit.

En fer la instal·lació prèvia per dur a terme la prova, ens vam trobar amb un sistema d'estructura interna molt sòlida destinat a la preservació d'objectes digitals que feia un èmfasi especial en l'organització i el sistema de preservació dels documents, i que cuidava més els aspectes estratègics i formals de la gestió dels objectes que no pas la seva presentació, amb conceptes que avui en dia ja són més normals, però que fa un any i mig eren avançats per a aquell temps.

La instal·lació dels components del sistema és habitual, és a dir, entorns *Linux/Unix* o *Windows*, amb requeriments de Java, Apache, Jakarta Tomcat, un gestor de bases de dades (*MySQL*, *PostgreSQL*, *Oracle* o la seva pròpia, *Mckoi*) i Ant per al desplegament dels serveis web. Ara bé, la manipulació d'aquests components estàndard i els propis del sistema el feien complicat d'entendre i d'ajustar en els seus paràmetres, i d'aprofitar-ne, així, tot el potencial.

A primer cop d'ull, un cop instal·lat i posat en marxa, cal destacar que el format de les metadades és XML amb l'estàndard de codificació i transmissió METS (http://www.loc.gov/standards/mets/).

Els serveis web, mitjançant WSDL (format XML que s'utilitza per a la descripció de serveis web) implementa dos tipus d'API (els API són un conjunt de rutines, protocols i eines per construir aplicacions): l'API d'administració (API-M) i l'API d'accés (API-A) a través de dos protocols: un "vell conegut", http get/post, i l'altre, SOAP (protocol de comunicació d'objectes entre diferents processos que intercanvien dades en format XML). Ambdós protocols són actius alhora. En la versió actual s'ha incrementat el nombre de serveis web.

Figura 5. Pàgina de consulta d'un dipòsit digital elaborat amb *Fedora*

Basant-nos en la versió 2 del document de l'Open Society Institute esmentada més amunt, agruparem els tests en tres nivells:

- Organització, compatibilitat i suport a diferents components
- Tractament de l'objecte digital (incorporació, metadades, objecte físic i conducta)
- Recuperació de l'objecte global

Cal destacar, com a punts favorables, la no-dependència d'una sola base de dades, el suport a diferents col·leccions, les diferents versions d'un objecte amb la garantia de la integritat de les dades preservades, i una estratègia ben definida i documentada de la preservació digital. També, quant al tractament de l'objecte digital, compost per un identificador persistent, metadades, *datastreams* (l'objecte o objectes digitals "físics" dipositats) i *disseminators* (mecanismes de conducta o, el que és el mateix, què pot fer el sistema per manipular l'objecte), s'incorporen amb aquest últim i per primera vegada en un gestor de continguts digitals de codi obert els mecanismes de conducta que permetien la manipulació i el tractament de les còpies que es visualitzen a la pàgina web. Per a imatges, s'inclouen dins del codi font crides a *MRSID* i *JPG2000* (programes de manipulació d'imatges), que proporcionaven zooms, retallades, escales de grisos, etc; en el cas dels textos, es proporcionaven conversions a l'instant (*on the fly*) a altres formats i manipulació d'OCR. L'avantatge evident és que únicament trobem aquests mecanismes en sistemes de gestió de continguts digitals de pagament.

Entre les mancances observades al moment oportú, podem destacar, pel que fa al tractament de l'objecte, l'existència de mecanismes complicats d'incorporació de dades i d'actualització, la manca d'un client amigable per a l'usuari autoritzat i les dificultats en la disseminació. Pel que fa a la recuperació, considerem un punt feble el fet de no tenir mecanismes eficients de cerca, truncament i recuperació de l'objecte o de les metadades.

Per finalitzar, la referència feta anteriorment a les API era imprescindible, ja que el motiu de la desestimació de *Fedora* pel grup es basa en aquesta condició: quan es proporcionen API vol dir que tenim un enorme potencial per desenvolupar, per programar. En el cas de *Fedora*, proporcionar les API volia dir absència de programació per a interfícies i clients. *Fedora* està concebut com una arquitectura sobre la qual desenvolupar sistemes d'usuari, no com un sistema d'usuari final i, per tant, calia construir els clients a tots els nivells. El grup, basant-se en els requeriments de l'encàrrec, buscava el contrari, un producte raonablement acabat. Aquest va ser un motiu suficient per descartar-lo.

Malgrat tot, algunes empreses comercials l'han adoptat com a base per desenvolupar els seus sistemes de gestió de continguts digitals; és el cas del producte *VITAL* de VTLS.

2.6 Projecte i-TOR

i-TOR (Tools and technology for Open Repositories), <http://www.i-tor.org/en>, va ser desenvolupat per IT-A (Innovative Technology-Applied), que és una secció del NIWI (Netherlands Institute for Scientific Information Services). Es tracta d'un conjunt d'eines per crear llocs webs, ja sigui un gestor de continguts, un portal, un dipòsit o d'altres similars. Això fa que sigui una eina molt flexible, però alhora fa pensar que, per implementar un dipòsit, caldrà fer tasques de programació. Quant als permisos dels usuaris, ens trobem amb la manca de limitació de l'accés a les diferents col·leccions segons els diferents usuaris. També en aquest cas, el nombre d'instal·lacions és molt baix i està restringit a Holanda. El grup va considerar oportú desestimar aquesta aplicació ja en la primera fase.

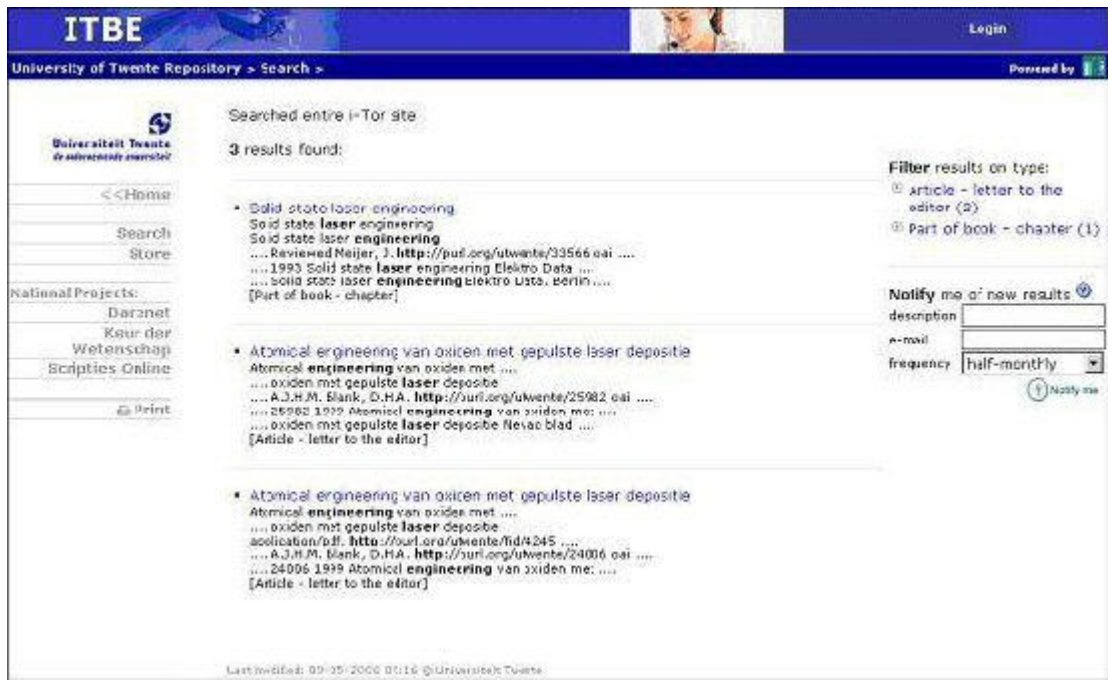

Figura 6. Pàgina de resultats d'una cerca al dipòsit de la Universitat de Twente, elaborat amb *i-Tor*

2.7 Projecte MyCoRe

MyCoRe (My Content Repositories), <http://www.mycore.de/engl>, és una evolució del programa *MILESS*, desenvolupat originàriament per la Universitat d'Essen. El programa estava pensat únicament per a la Universitat d'Essen i no s'ajustava a les necessitats d'altres institucions. Aleshores, les universitats de Jena i Leipzig van crear *MyCoRe*, que és un "nucli" altament configurable per implementar dipòsits basats en Java i XML. El fet que sigui només un "nucli" i no una aplicació acabada que es pugui instal·lar i fer funcionar comporta que calgui una gran tasca de programació al darrere, ja que no hi estan implementats ni la gestió d'usuaris i contrasenyes, ni la gestió de submissions ni la generació de llistats i estadístiques. Tampoc no disposa de cap estratègia de preservació de la informació digital. Per acabar, el fet que el nombre d'instal·lacions fos molt baix i que estigués restringit a Alemanya i Suècia van motivar que desestiméssim aquesta aplicació ja en la primera fase.

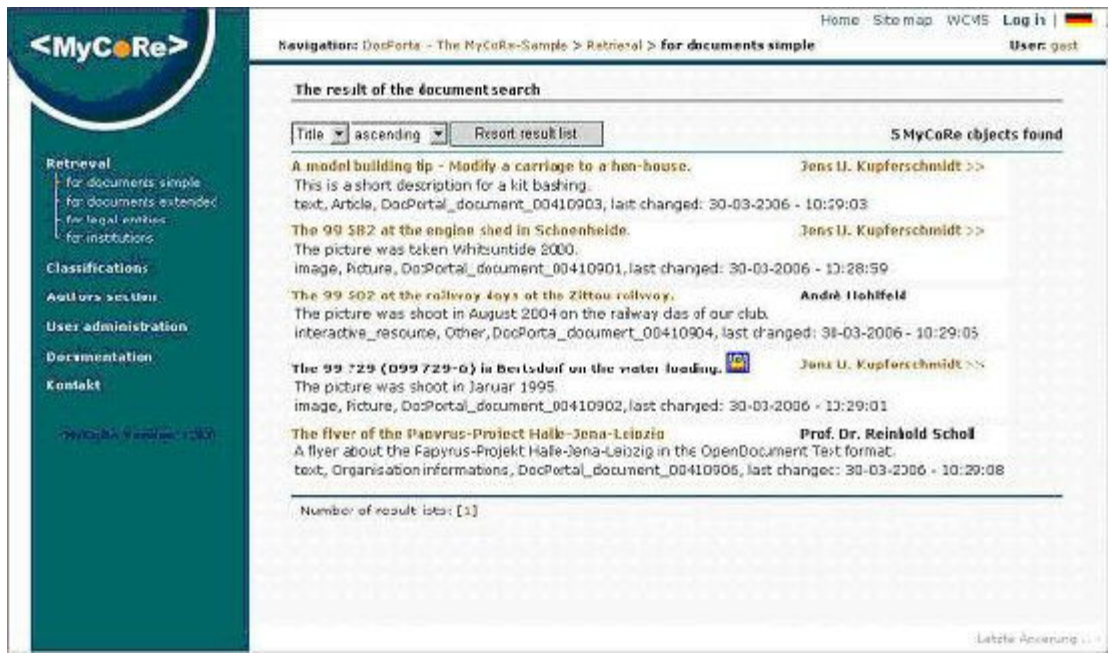

Figura 7. Pàgina de resultats d'una cerca simple amb *MyCoRe*

3 Conclusions

3.1 Sistema seleccionat

Després de força reunions i sessions de treball en grup, es va considerar que els tres sistemes que havien arribat al final del procés de selecció cobrien les necessitats demanades i estaven a l'altura tècnica i d'implantació de qualsevol altre sistema comercial, però també es va estimar que, donats els objectius perseguits i la situació en aquell moment, *DSpace* reunia les condicions òptimes. Les raons més destacables per prendre aquesta decisió a favor de *DSpace* van ser el grau d'acabat del producte, l'important nombre d'instal·lacions en diferents i prestigioses universitats del món i la bona interfície d'usuari i d'administrador. Tot això permetia al CBUC donar una solució tècnica immediata per a la creació del dipòsit. D'altra banda, també es van destacar alguns punts febles, entre els quals: la manca de suport a la cerca amb diacrítics i el baix coneixement de la plataforma tecnològica per part de les biblioteques del CBUC (plataformes *Java* i *Lucene*).

Així, doncs, el grup va adreçar a la Comissió Tècnica del CBUC la proposta d'usar *DSpace*, que va ser aprovada en la reunió següent, el maig de 2004.

3.2 Implantació

Des de la decisió d'usar *DSpace* com a programari per gestionar els dipòsits dins del CBUC, s'han creat diferents grups de treball i ja hi ha actius i en fase de preparació diferents dipòsits. Com que un dels objectius del procés era crear una comunitat d'usuaris del producte de manera que tinguéssim un suport local i que els desenvolupaments propis es poguessin aprofitar en altres institucions, es va decidir crear tres grups estables:

- Grup de sistema. Tracta els temes més informàtics del producte. Arran del treball d'aquest grup, es va poder resoldre la cerca amb diacrítics. Posteriorment, aquest problema ja ha estat resolt en el cercador intern de *DSpace, Lucene*.
- Grup de metadades. Tracta, adapta i recomana els diferents conjunts de metadades (format Dublin Core en el cas de *DSpace*) que haurien d'usar les institucions en els seus dipòsits i en el del CBUC.
- Grup de gestió. Tracta els temes d'administració i ús del programa.

El resultat pràctic de la tria d'aquest sistema és Recercat, <http://www.recercat.net>, un dipòsit i recol·lector de metadades cooperatiu del CBUC que inclou la literatura grisa de recerca de les universitats i dels centres d'investigació de Catalunya, com ara articles encara no publicats (*preprints*), comunicacions en congressos, informes de recerca, *working papers*, projectes de final de carrera, memòries tècniques, etc. En una primera etapa de definició del que havia de ser aquest dipòsit es va proposar d'incloure-hi només *working papers*. Posteriorment es va constar la necessitat d'ampliar l'abast del dipòsit a la literatura grisa de recerca més en general.

A més, diferents universitats i altres institucions estan utilitzant *DSpace* per als seus dipòsits locals i més específics. Destaquem, especialment, el cas de la Universitat de Girona, amb un dipòsit de material audiovisual, <http://diobma.udg.es:8080/dspace/index.jsp>, enllaçat amb un servidor de vídeo a demanda dels usuaris. La Universitat Politècnica de Catalunya (UPC) també ha posat en marxa diferents dipòsits institucionals basats en aquesta aplicació: *DSpace.Revistes UPC*, <http://e-revistes.upc.edu>, que permet l'accés obert als articles de les revistes publicades per les unitats i pels grups de recerca, i *DSpace.E-prints UPC*, <http://e-prints.upc.edu>, que facilita la publicació en accés obert dels treballs de recerca de la Universitat, a més del portal *UPCommons*, <http://upcommons.upc.edu>, que dóna un accés unificat als continguts dels diferents dipòsits que el Servei de Biblioteques i Documentació ha desenvolupat, mitjançant la implementació d'un recol·lector de metadades.

Data de recepció: 15/04/2006. Data d'acceptació: 13/05/2006.

Notes

[1] Open Archives Inititiative-Metadata Harvesting Protocol, <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[2] "A guide to institutional repository software". V. 3.0. New York: Open Society Institute, January 2004. <http://www.soros.org/openaccess/software>. [Consulta: 29/04/2006].

[3] European Organization for Nuclear Research, <http://www.cern.ch/>, és el centre europeu de recerca en física i és el lloc on es va crear el World Wide Web.

[4] La gran quantitat de paquets "externs" fa que sigui molt més recomanable la instal·lació en *Linux*, en què la majoria d'aquestes aplicacions ja s'acostumen a trobar precompilades, que no pas en sistemes *Unix* propietaris, com ara *Solaris*, en què la feina d'instal·lació i, sobretot, de manteniment d'aquests requisits augmenta massa. Vegeu <http://cdsware.cern.ch/download/INSTALL>. [Consulta: 03/05/2006].

[5] "UTF-8 (*8-bit Unicode Transformation Format*) és una codificació de caràcters de longitud variable per a Unicode que utilitza grups de

bytes per representar l'estàndard d'Unicode per als alfabets de molts llenguatges del món". <http://en.wikipedia.org/wiki/UTF-8>. [Consulta: 03/05/2006].

[6] Per exemple, es pot definir la col·lecció *Vídeos* com tots els registres que tenen el valor *VIDEOS* al subcamp 980$a, o *Fotografies de la Guerra Civil* per als que tinguin el valor *FGC*.

[7] *HOWTO MARC Your Bibliographic Data*, <http://cdsweb.cern.ch/admin/howto/marc.html>. [Consulta: 03/05/2006].

[8] El que habitualment s'anomena *self-archiving*.

[9] <http://archives.eprints.org/?action=browse>.

[10] *Citebase Search*, <http://www.citebase.org/>.

[11] *Registry of Open Access Repositories*, <http://archives.eprints.org/>.