Selecció i implementació d'un motor de cerca a la Biblioteca Virtual de la UOC

Autors:

Jordi Serrano Muñoz. jserrano@uoc.edu
Responsable d'avaluació i explotació dels serveis bibliotecaris

Adoració Pérez Alarcón. dora@uoc.edu
Directora de la Biblioteca Virtual

- Resum
- Introducció
- Unes dades preocupants
- El procés de selecció
- Els productes
- Instal·lació
- Definició de les col·leccions
- Les parts més fàcils
- Una mica de complexitat
- DinaWEB/DIMAX
- Sumaris i articles de revistes
- La part més difícil
- Disseny de la interfície i restriccions d'accés
- Conclusions
- Bibliografia

RESUM

A partir de diferents enquestes de satisfacció institucional, i de l'anàlisi de l'arxiu "log" del servidor biblioteca respecte a l'ús i el comportament dels usuaris es va detectar que cada cop era mes complex accedir als continguts i serveis de forma proporcional al creixement d'aquests i del nombre d'usuaris.

El creixement de recursos i dels diferents aplicatius desenvolupats a la Biblioteca Virtual de la UOC (BUOC), va fer necessari la selecció i implementació d'un motor de cerca que facilites de forma global l'accés als recursos d'informació i serveis ofertats a la comunitat virtual de la UOC en funció de la tipologia d'usuari, l'idioma i el seu entorn d'aprenentatge.

En aquest article, s'exposa el procés d'anàlisi de diferents productes, i la implementació de Verity a la BUOC amb els desenvolupaments realitzats als diferents aplicatius perquè el motor de cerca pugui realitzar la seva funció.

INTRODUCCIÓ

Des dels seus inicis al curs 1995-1996, la BUOC té com a missió donar suport a la activitat docent, de recerca i de gestió a la comunitat virtual de la Universitat Oberta de Catalunya (UOC) per mitjà del seu Campus Virtual.

El Campus Virtual, es bàsicament una Intranet educativa on interactuen els membres de la UOC, i des d'on la BUOC dona servei a aquests membres per mitjà del seu servidor web.

Durant el curs 1995-1996, s'oferia als usuaris l'accés per web a una base de dades Oracle que contenia la informació bibliogràfica dels documents físics disponibles i una bústia electrònica d'atenció a l'usuari, mentre es treballava de cara al curs 1996-97 en la instal·lació de VTLS i l'adaptació de la passarel·la web per incloure la demanda de préstecs en línia (1) així com en la creació de formularis que actuessin com a taulell virtual d'informació i es seleccionaven recursos digitals accessibles per mitjà de pàgines estàtiques.

A partir del curs 1996-1997, es van incorporant documents electrònics de compra i també documents generats per la universitat i per la pròpia biblioteca i s'inicia com la digitalització de sumaris i resums de llibres físics i dels sumaris de les revistes en paper subscrites per la UOC. En el curs 1999-2000 se substitueixen les pàgines estàtiques amb recursos d'informació per un gestor de recursos digitals conegut per DinaWEB (2) amb estàndard Dublin Core, del que actualment disposem d'una nova versió (DIMAX) que inclou a més els estàndards de e-learning IMS i SCORM.

En paral·lel, el nombre d'usuaris passa d'aproximadament de 200 als casi 26.000 del curs actual i la oferta d'estudis s'amplia a diferents iniciatives com Màsters, Postgraus, accés a majors de 25 anys, Escola virtual d'Idiomes i un Doctorat en Societat de la Informació.

UNES DADES PREOCUPANTS

Cada final de semestre, la UOC realitza una enquesta entre els alumnes, les preguntes relatives a la BUOC mostraven un grau de satisfacció elevat però a les respostes lliures anaven sorgint comentaris relatius a la complexitat de recuperar la informació adient entre tot el conjunt d'informació disponible a la Biblioteca, així com l'estudi del "log" ens advertia de que cada vegada els usuaris passaven més temps a la biblioteca i que les passes per arribar a la informació es feien més complexes.

Aquests aspectes, tenien una correlació directe amb l'increment de recursos digitals com a conseqüència de la oferta formativa, que s'intenta minimitzar amb la implementació de prestatgeries virtuals o de petites biblioteques amb el material bàsic per a cada assignatura i en paral·lel s'anaven incrementat els serveis especialitzats amb els diferents aplicatius per a facilitar-ne l'accés.

L'any 2000, l'usuari disposava de l'OPAC per accedir a documents físics i digitals disponibles a la UOC o en el servidor de la biblioteca, del DinaWEB, per accedir a recursos digitals tant interns com externs, i també un petit motor de cerca que permetia accedir als documents del servidor i als sumaris i resums, així com als sumaris de les publicacions periòdiques, a més de serveis com els de notícies i de distribució electrònica de sumaris.

Per tant, l'usuari es trobava, des de la pàgina principal del servidor web de la biblioteca, amb la necessitat de prendre una decisió prèvia respecte al tipus de material que volia recuperar (paper, digital, digital creat per la Biblioteca, etc.) que l'obligava, en la majoria dels casos, a repetir la cerca en els diferents aplicatius.

Aquest escenari, ens va dur a un procés de reflexió en el que es va veure la necessitat de crear un únic punt d'accés, amb independència dels aplicatius existents i futurs, que millorés la recuperació d'informació en el servidor web adaptant-se als diferents formats i dels aplicatius.

En paral·lel al desenvolupament i implementació del motor de cerca, un test d'usuabilitat realitzat en el Campus virtual i a la Biblioteca confirmava la complexitat d'aquesta a l'hora d'utilitzar-la (3), i ens va confirmar aquesta necessitat a més de redissenyar la interfície web de la biblioteca virtual.

EL PROCÉS DE SELECCIÓ

A part del volum d'informació disponible, ens trobàvem enfront de diferents estàndards com MARC, XML, IMS, SCORM i W3C i de diferents formats de documents: MS-Office, RTF, HTML, PDF i ASCII entre altres

Els aspectes anteriors, feien necessari localitzar una eina que facilités l'accés a la informació independentment del suport, format, entorn d'usuari i idioma de l'entorn i dels recursos.

En la definició del producte, es requeria que fos de fàcil administració, que les cerques es fessin de forma familiar a l'usuari (tipus Google, Altavista, ...), que la visualització de resultats tingués una forma clara i parametritzable i que contemplés diferents tipus de cerca.

Es va redactar un documents de requeriments (4) que contemplava tot un conjunt d'aspectes agrupats en cinc grans apartats:

1. Hardware
   - Plataforma SUN Solaris
   - RAM
   - Compressió índexs (espai de disc)

2. Indexació per
   - Arxiu
   - Directori
   - Arxius remots a partir de la URL
   - Actualitzable de forma automàtica dels canvis en els continguts de la pàgina

3. Formats

- HTML
- Text
- PDF
- MS-Office
    - XLS (Excel)
    - DOC (Word)
    - MDB (Access)
    - PPT (PowerPoint)
- Missatges de correu electrònic

4. Tipus de Cerca
    - Sinònims
    - Paraula
    - Camps
    - Rang (>,<,=)
    - Frase i proximitat
    - Booleana
    - Caràcters de substitució (?,*,..)
    - Llenguatge natural
    - Semàntica
    - Query by Example

5. Part client (usuari)
    - Personalització de la cerca
    - Guardar/actualitzar cerques (perfils)
    - Push

6. Multifunció
    - Catàleg
    - Web
    - Internet
    - Motor/s de cerca
    - DinaWEB
    - Bases de dades remotes

El punt 6 referit a la multifunció, es el que es va considerar més rellevant, l'eina, tenia que dialogar amb els aplicatius existents a la BUOC.

A partir dels punts anteriors, es va fer un treball de cerca bibliogràfica per a familiaritzar-se amb el tema entre els que cal destacar els butlletins digitals de Search Engine Watch (5) i Axandra Search Engine Facts (6)

ELS PRODUCTES

Si ens movem en el camp dels motors de cerca més complexos que els tradicionals i que incorporaven més prestacions i possibilitats ens trobaven en un mercat en expansió i en constant creixement . Al 1997, una tercera part del mercat d'aquest productes es centrava en cinc empreses i productes (7):

- PC Docs Fulcrum Search Server de Fulcrum Technologies:
- BASIS de Information Dimensions:
- Dataware II knowlegde de Dataware Technologies
- Verity: Information Server de Verity Inc.
- Excalibur Retrievalware de Excalibur Technologies

Altres autors a la mateixa època els reduïen a tres (8):

- Excalibur Retrievalware
- Verity
- PC Docs

Durant el 1998 i 1999, es van incorporar al mercat noves aplicacions que també calia considerar:

- INMAGIC DB/TEXT Intranet Spider de Inmagic Inc.
- Ultraseek Server de Infoseek Technology
- Personal Librarian Web Turbo de America Online
- InfoMagnet de CompassWare Development
- Altavista Search Intranet de Altavista Inc.

Es va fer un estudi de tots els productes i el els casos en que va ser possible en va contactar amb els proveïdors per a una demostració, a més de provar-ho en llocs en explotació.

Els resultats es van incloure en taules com la d'exemple:

Producte	PC Docs	Basis	Dataware II knowledge	Verity Information Server	Excalibur Retrieval Ware 6.0	InfoMagnet	Ultraseek Server	PLWeb Turbo	Inmagic DBText Intranet Spider	Altavista Search Intranet
Paraula	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
Frase/Proximitat	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
Proximitat	Si	Si	Si	Si	Si	Si	No	Si	Si	Si
Booleana	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
Rang	Si	Si	Si	Si	Si	No	No	Si	Si	Si
Caràcters de substitució	Si	Si	Si	Si	Si	No	No	Si	Si	Si
Llenguatge natural	Si	No	?	Si	Si	Si	Si	Si	No	Si
Sinònims	Si	No	?	Si	Si	Si	Si	Si	No	Si
Semàntic	Si	No	?	No	Si	No	No	No	No	No
Query by Example	Si	No	?	Si	Si	Si	No	Si	No	No

Figura 1. Taula comparativa dels diferents productes relativa als tipus de cerca

Ponderats els resultats, les nostres opcions es varen decantar per Excalibur Retrievalware i Verity, i finalment ens vam decidir per aquest darrer no tant sols pel cost, sinó per el suport tècnic que contemplava a una persona amb experiència en el món de les biblioteques.

INSTAL·LACIÓ

L'any 2002, es va iniciar la instal·lació de Verity (Verity Information Server 3.7) en el servidor de la BUOC amb les següents característiques de maquinari: Sun Ultra-60 amb processador 2 UltraSPARC-II a 296 MHz i una memòria de 512 MB amb el sistema operatiu Solaris 8 i amb la versió 1.3.26 de Apache

DEFINICIÓ DE LES COL·LECCIONS

Les col·leccions s'entenen a Verity com a grups de documents. Donades les característiques de la BUOC es va procedir a agrupar la informació i els documents en funció del idioma:

	Català	Castellà	Anglès	Comú (Informació o documents independents de l'idioma o entorn d'usuari)
Documents electrònics				X
OPAC	X	X	X	
Recursos digitals DinaWEB/ DIMAX	X	X	X	
Sumaris i resums de llibres				X
Articles dels sumaris de revistes UOC				X
Pàgines estàtiques amb informació i serveis	X	X	X	

Taula 1. Col·leccions creades en funció de l'idioma

Indexades aquestes col·leccions, i realitzades les primeres proves de cerca i visualització, es va decidir subdividir addicionalment la col·lecció de documents electrònics per format , ja que en donava més flexibilitat a l'hora de visualitzar els resultats.

LES PARTS MÉS FÀCILS

Les parts més fàcils a l'hora de l'indexació, van ser els documents electrònics i les pàgines d'informació i serveis. El procés es senzill com a qualsevol motor de cerca i la recuperació mostra la informació inclosa a les metadades de la pàgina o del document: títol, autor i descripció. que en cas de no tenir ens mostra la descripció que genera el propi cercador.

Es van aplicar per mitjà del llenguatge de Verity (SearchScipt) especialment quan l'autor era la pròpia biblioteca de la UOC o la metadada del Títol en els documents electrònics de compra, es en blanc o conté text del tipus "file c://mydocuments/...."

Casi per casualitat, les divisió en col·leccions va permetre afegir identificadors dels tipus “Document electrònic”, “Informació i serveis” o “Manual d’ús” previs al resultat de la cerca

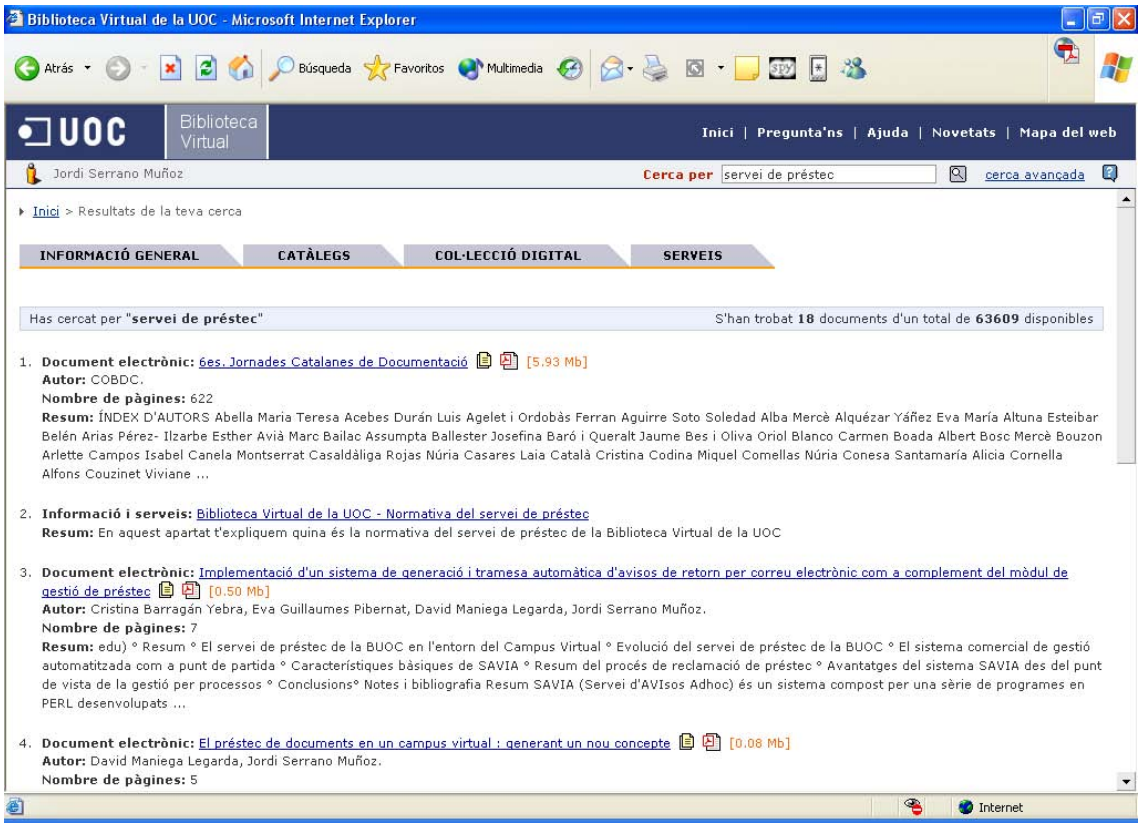

Figura 2. Exemple de resultats de documents electrònics i pàgines d’informació i serveis

UNA MICA DE COMPLEXITAT

SUMARIS I RESUMS

El següent repte van ser els sumaris i els resums dels llibres físics disponibles a la Universitat. La indexació no revestia més complexitat que en el cas anterior, però la visualització no aportava més valor.

Es va decidir incloure l’accés a l’OPAC amb la informació bibliogràfica i la possibilitat de demanar en préstec el document objecte del resum o del sumari. La solució va partir del nom del document que està relacionat amb l’etiqueta 035 de VTLS: el SearchScript permet “trossejar” el nom de l’arxiu i “enganxar-ho” a una crida predefinida a VTLS.

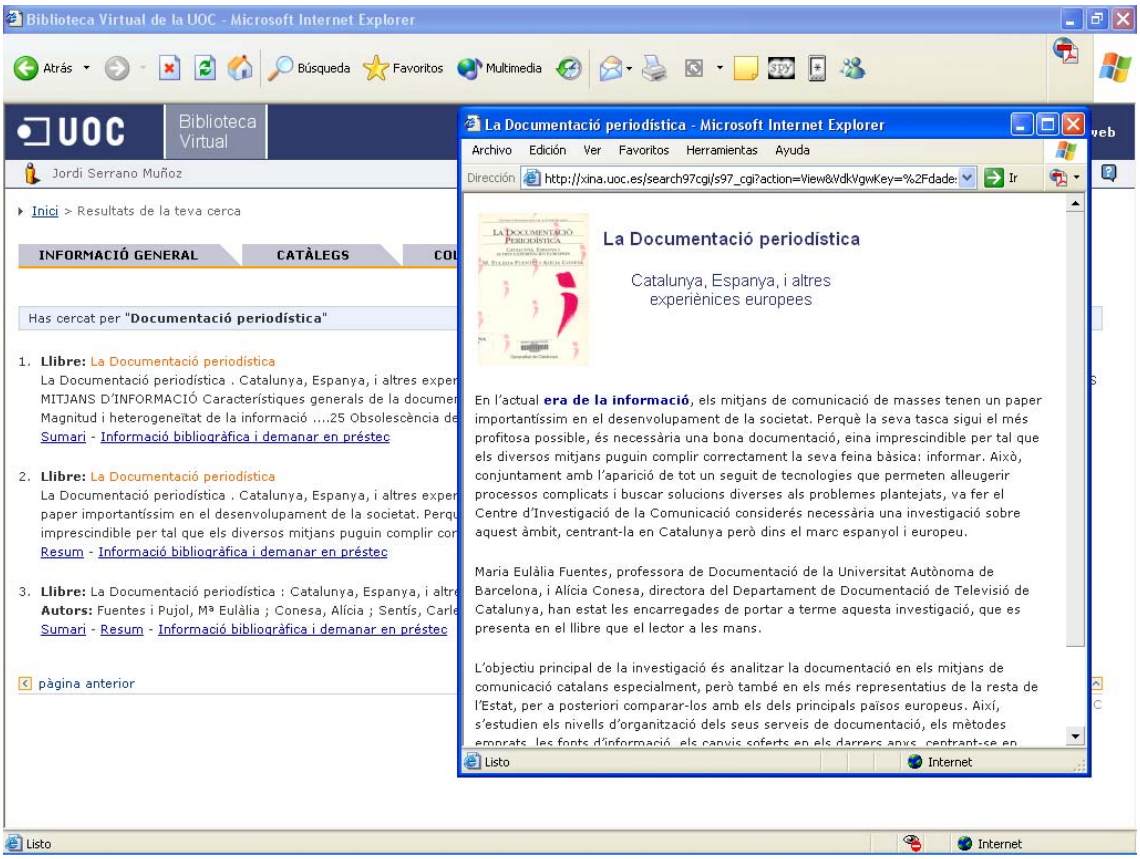

Figura 3. Sumari i resum d'un document amb accés a l'OPAC i al préstec

DinaWEB/DIMAX

Dins el gestor de recursos digitals els primers intents d'indexació, es van realitzar directament contra la base de dades, però el resultats no van ser prou satisfactoris. Es va decidir fer una extracció etiquetada en XML i definir els camps dins els paràmetres de la col·lecció.

Cada 24 hores es realitza, de forma automatitzada una extracció de l'arxiu XML que s'exporta al servidor web de la biblioteca, on Verity actualitza la informació continguda a l'arxiu XML de cada idioma:

Exemple de recurs en XML:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ROWDATA>
        <ROW>
                <ID_RECURS>10</ID_RECURS>
                <DATA>27/08/99</DATA>
                <DATA_REVISIO>18/11/03</DATA_REVISIO>
                <DATA_CADUCITAT></DATA_CADUCITAT>
                <CONTROL_TREN></CONTROL_TREN>
                <USUARI></USUARI>
                <PWD></PWD>
                <AUTOR></AUTOR>
```

```
            <PROVEIDOR></PROVEIDOR>
            <PUBLICAT_A></PUBLICAT_A>
            <PAIS>Espanya</PAIS>
            <IDIOMA_RECURS>es</IDIOMA_RECURS>
            <FORMAT>HTML</FORMAT>
            <PROTOCOL>HTTP</PROTOCOL>
            <TEXT_COMPLERT>0</TEXT_COMPLERT>
            <AMBIT_GEOGRAFIC>Espanya</AMBIT_GEOGRAFIC>
            <TITOL>EDIT. Base de datos de editoriales españolas e hispanoamericanas</TITOL>
            <ABSTRACT>Base de dades del Ministeri d'Educació i Cultura que inclou un cens d'editorials
            espanyoles i les més significatives d'hispanoamericanes.</ABSTRACT>
            <URL>http://www.mcu.es/bases/spa/edit/EDIT.html</URL>
            <MANUAL></MANUAL>
            <MANUAL_UOC></MANUAL_UOC>
            <BUIDAT_A></BUIDAT_A>
            <PERIODICITAT></PERIODICITAT>
            <ISSN_ISBN></ISSN_ISBN>
            <EDICIO></EDICIO>
            <MATERIA_CAT>Editors i edició - </MATERIA_CAT>
            <MATERIA_CAS>Editores y edición - </MATERIA_CAS>
            <MATERIA_ANG>Publishers and Publishing - </MATERIA_ANG>
            <LLISTA_NODES>Editorials(663) -Documentació(268) -</LLISTA_NODES>
        </ROW>
</ROWDATA>
```

SUMARIS I ARTICLES DE REVISTES

A partir de l'experiència anterior, va ser més fàcil treballar aquest conjunt de documents, ja que la gestió dels sumaris i els seus articles comparteixen una estructura similar al DinaWEB/DIMAX. El procediment va ser el mateix: extreure un arxiu etiquetat en XML.

Però com en els sumaris i resums dels llibres, la informació no aportava cap valor afegit, i es va veure necessari enllaçar-ho amb el sistema de Distribució electrònica de sumaris (DESU) que incorpora la possibilitat de fer comandes al Servei d'Obtenció de Documents (SOD).

Quan l'usuari recupera un article, Verity mostra les dades: títol, autor, paginació, revista, volum, any, etc.. i facilita l'accés a l'índex de la revista, al sumari del número i a la funcionalitat de demanar una còpia de l'article (Figura 4).

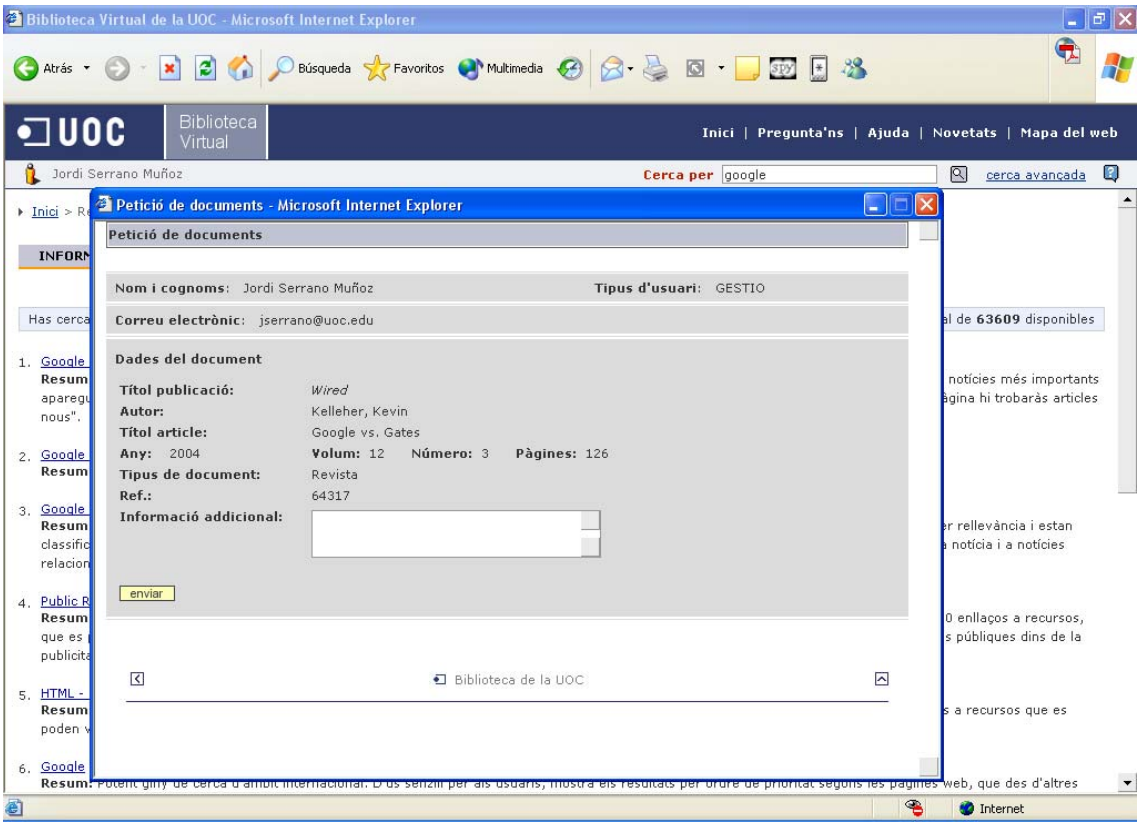

Figura 4. Comanda al Servei d'Obtenció de Documents (SOD) a partir de Verity

LA PART MÉS DIFÍCIL

La part més difícil, va ser la indexació de l'OPAC, ja que indexava, tot un conjunt d'elements que no eren rellevants per a l'usuari: les pantalles intermitges, informació dels exemplars, etc.

El que interessava, era indexar les pantalles d'informació bibliogràfica (Card Screen) i es va optar per fer un petit programa en Perl a VTLS, que generava una pàgina estàtica a partir de l'etiqueta 035 amb la crida a cada pantalla de la informació bibliogràfica. Es donaven instruccions a Verity d'indexar totes les URLs de la pàgina estàtica

A partir d'aquí es va detectar que el robot o "spyder", anava saltant per tots els enllaços de cada pantalla intermitja pel que es va afegir a la passarel·la web on genera la pàgina, les metadades relatives als robots i motors de cerca de indexar tant sols la pàgina recuperada i no continuar indexant els enllaços inclosos:

<meta name="robots" content="index, nofollow">

Resolt aquest aspecte, es va comprovar que no era gaire amigable la visualització dels resultats, pel que es va decidir modificar novament la passarel·la web i afegir en PERL instruccions que etiquetessin alguns dels camps de VTLS: autor/s, títol/s, matèries en català i castellà, edició, dades editorials i en el cas de bibliografia recomanada i materials didàctics, la assignatura o assignatures a les que està adscrit el document..

Aquest aspecte, va ajudar a aportar valor a la informació bibliogràfica dels documents, ja que l'usuari a més de la informació aportada per l'OPAC, pot accedir al sumari i al resum, disposar de la possibilitat de demanar el document en préstec, i en el cas de la bibliografia recomanada, es facilita l'accés al llistat de tots els documents relacionats amb l'assignatura (figura 5) .

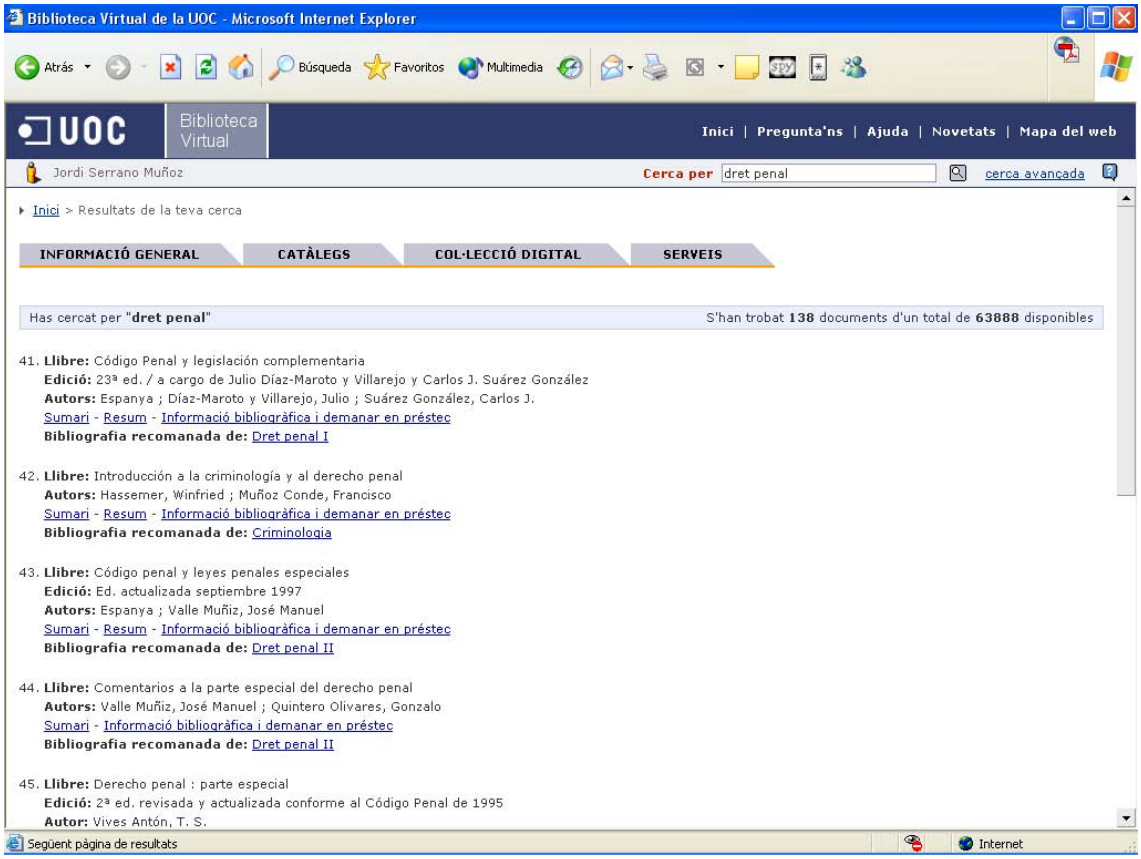

Figura 5. Documents de l'OPAC amb accés a la informació complementària (sumari, resum), sol·licitud de préstec i a la bibliografia de l'assignatura

DISSENY DE LA INTERFÍCIE I RESTRICCIONS D'ACCÉS

Un cop testejats i validats els diferents processos d'indexació i recuperació, es va procedir a dissenyar de la interfície de consulta i de resultats per mitjà de plantilles i a configurar l'opció de cerca avançada.

Es van tenir en compte els resultats i les recomanacions de l'auditoria d'usabilitat realitzada a tota la biblioteca virtual incloent-hi les recomanacions del W3C respecte a persones amb discapacitats visuals.

Finalment, un darrer aspecte, probablement un dels més rellevants de la implementació, va ser la restricció d'accés en relació a les llicències o els serveis a parts dels continguts de la BUOC.

Fins aleshores, el propis aplicatius DinaWEB/DIMAX , DESU o distribució de sumaris i l'OPAC, eren els que aplicaven aquesta restricció. Bàsicament es tractava de configuracions predefinides de tipus d'usuari dins del Campus Virtual que certificaven o autoritzaven l'accés als recursos, a documents protegits o als serveis, per mitjà d'una passarel·la transparent per a l'usuari .coneguda dins la UOC com a TREN.

L'esquema d'aquesta passarel·la (Figura 6), es a partir del navegador de l'usuari i de la connexió a algun dels "Frontends" del Campus Virtual, que llegeix en el servidor d'aplicacions, els "privilegis" dels usuaris com aules d'estudi, expedient acadèmic i en el nostre cas les funcionalitats de la biblioteca accessibles.

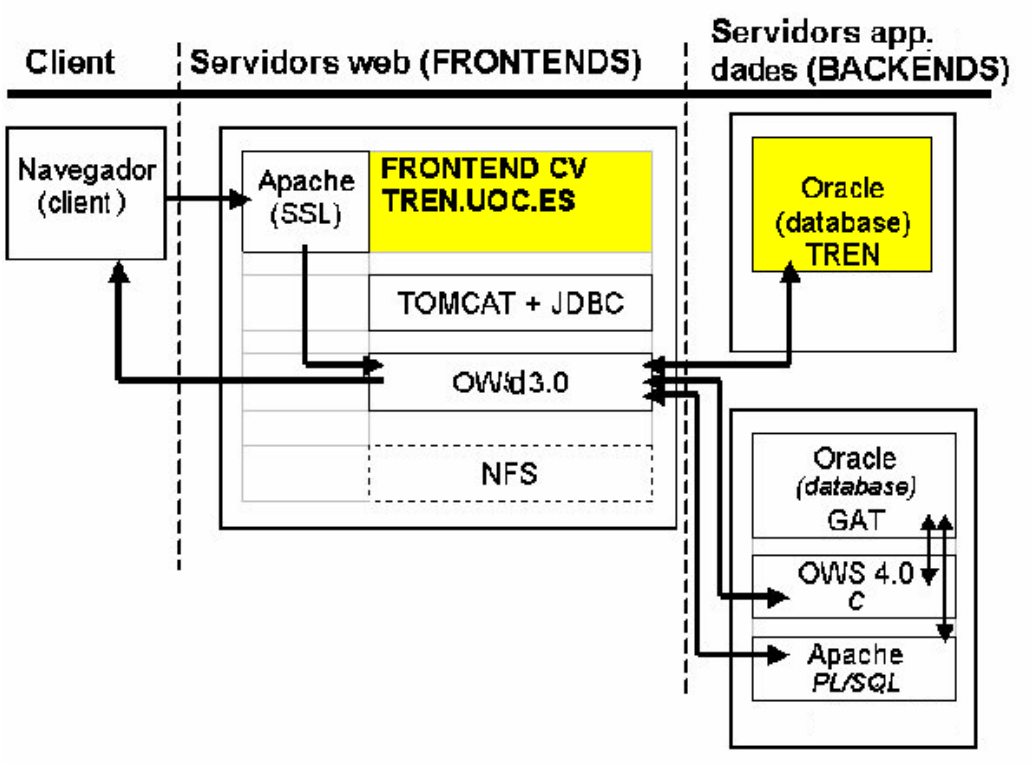

Figura 6. Esquema de la passarel·la TREN dins el Campus Virtual

Aprofitant la nova versió de la biblioteca virtual, es va optar per desenvolupar un aplicatiu que des de la biblioteca "agafa" durant tota la connexió al Campus la identificació i el codi de la sessió de connexió de l'usuari i es va "arrossegant" o "transportant" en cada una de les accions que l'usuari realitza dins del servidor web.

Aquesta funcionalitat es va incloure com una variable dins dels paràmetres de Verity, i així en una connexió externa al Campus Virtual des d'internet al servidor de la BUOC (http://biblioteca.uoc.edu) aquesta variable es buida, i per tant no permet l'accés als recursos digitals de pagament i a serveis com el préstec o de serveis documentals de subscripció.

En el Campus virtual amb aquesta informació inclosa dins de la variable de Verity, permet l'accés exclusiu, en funció de l'entorn d'aprenentatge de l'usuari, als continguts i serveis propis de la biblioteca, seleccionats i elaborats per a cada un d'aquests entorns.

CONCLUSIONS

Donat el creixement d'usuaris, de recursos, d'aplicatius i de serveis, des de la BUOC, es va veure necessari implementar una eina que facilités l'accés a tots ells. La idea original era la implementació d'un motor de cerca similar als existents a internet, però una vegada analitzats, es va veure la oportunitat d'afegir valor als resultats de les diferents cerques i no limitar-se a mostra una llista de resultats mes o menys adients.

A l'hora de seleccionar l'eina, es van prioritzar requeriments que permetessin interactuar o dialogar amb els aplicatius de continguts (DinaWEB/DIMAX i OPAC) i de serveis (préstec, SOD, serveis a mida i sumaris) i altres aplicatius que es poguessin desenvolupar en un futur.

La complexitat en la tasca d'implementació, va ser originada pel diàleg amb els aplicatius, però la flexibilitat del llenguatge (SearchScript) que incorpora Verity i el desenvolupament de petites aplicacions, van permetre superar aquestes dificultats.

La coincidència amb l'auditoria d'usabilitat, va permetre adaptar el producte a les recomanacions d'aquesta auditoria i incloure-ho amb la nova versió de la biblioteca virtual.

El resultat ha estat un nou punt d'accés que engloba totalment el conjunt de recursos i serveis disponibles a la UOC, sense excloure les opcions tradicionals que fins el moment de la implementació s'ofertaven.

BIBLIOGRAFIA

David Maniega Legarda, Jordi Serrano Muñoz (1997). *El préstec de documents en un campus virtual: generant un nou concepte*. A: 6es Jornades Catalanes de Documentació. Barcelona, 23-25 octubre 1997. <http://biblioteca.uoc.edu/cgi-bin/pass/byteserver.pl/docs_elec/ponencies/1909.pdf> [Consulta 01/04/04]

Jordi Serrano Muñoz, David Maniega Legarda, Cristina Barragán Yebra, Juanjo Martí Manzano, Clara Beleña, Ramon Capillas (1999). *DinaWEB: l'organització de recursos accessibles en línia a la Biblioteca Virtual de la Universitat Oberta de Catalunya*. A: 7es Jornades Catalanes de Documentació. Barcelona, 4-7 novembre 1999 <http://biblioteca.uoc.edu/cgi-bin/pass/byteserver.pl/docs_elec/ponencies/3525.pdf> [Consulta 01/04/04]

David Maniega Legarda (2002). *Aplicación de un estudio de usabilidad en bibliotecas digitales: la Biblioteca Virtual de la UOC*. A: Workshop CALSI - Universitat de València, 22-23 octubre 2002. <http://biblioteca.uoc.edu/cgi-bin/pass/byteserver.pl/docs_elec/comunicacions/12882.pdf> [Consulta 01/04/04]

Biblioteca de la UOC (1999). *Selecció d'un motor de cerca per la Biblioteca de la UOC: Anàlisi de productes*. (Document intern)

Search Engine Watch: <http://searchenginewatch.com/> [Consulta 01/04/04]

Axandra Search Engine Facts: <http://www.axandra.com/> [Consulta 01/04/04]

Chris Nerney (1997). *Searching for true knowlegde*. Network World. Vol 14(24) p.42 (june 16, 1997)

Robert Boeri. Martin Hensel (1997). What's next for text: Retrieval trends past, present and push. EMedia Professional. April 1997. p. 53