



Repositorios colectivos de e-información

Ricard de la Vega, Marc Boix, Joan Cambras, Natalia Torres, M^a Teresa Novo, Jordi Prats, Núria Comellas, Sandra Reoyo, Ramon Ros • 16-11-06

Resumen

Desde 1999, el Centre de Supercomputació de Catalunya (CESCA)^[1] promociona actividades para poner en acceso abierto contenido en internet, contribuyendo al avance de la e-Ciencia en nuestro país.

Junto con el Consorci de Biblioteques Universitàries de Catalunya (CBUC)^[2] se han creado tres repositorios: Tesis Doctorales en Red (TDR)^[3], Dipòsit de la Recerca de Catalunya (RECERCAT)^[4] y Revistes Catalanes amb Accés Obert (RACO)^[5]. En septiembre de 2006, la Biblioteca Nacional de Catalunya^[6] ha puesto en marcha otro ambicioso repositorio en colaboración con el CESCA, Patrimoni Digital de Catalunya (PADICAT)^[7].

Palabras clave: repositorio, e-ciencia, investigación, acceso abierto, archivo web

Summary

Since 1999, the Centre de Supercomputació de Catalunya (CESCA) promotes activities to put internet content in open access, so it's contributing to the e-Science progress of our country.

CESCA together with Consorci de Biblioteques Universitàries de Catalunya (CBUC) have created three repositories: Tesis Doctorales en Red (TDR), Dipòsit de la Recerca de Catalunya (RECERCAT) and Revistes Catalanes amb Accés Obert (RACO). In September of 2006, the Biblioteca Nacional de Catalunya has started another ambitious repository in cooperation with CESCA, the Patrimoni Digital de Catalunya (PADICAT).

Keywords: repository, e-science, research, open access, web archiving

Introducción

El Scholarly Publishing and Academic Resources Coalition (SPARC)^[8] define los repositorios electrónicos institucionales como una colección digital que captura y preserva los resultados intelectuales de una o más de una institución. Se podrían destacar las siguientes características:

- Contienen documentos generados por las instituciones.
- De contenido académico.
- Con voluntad acumulativa y persistente.
- De manera abierta e interoperable.

El Centre de Supercomputació de Catalunya (CESCA), junto con el Consorci de Biblioteques Universitàries de Catalunya (CBUC), ha puesto en marcha tres repositorios electrónicos colectivos institucionales de e-información: TDR, RECERCAT y RACO.

TDR (Tesis Doctorales en Red) es un repositorio que contiene en acceso abierto tesis doctorales presentadas en quince universidades españolas. RECERCAT (Dipòsit de la Recerca de Catalunya) dispone de literatura gris de universidades y centros de investigación catalanes, como artículos aún no publicados (*preprints*), comunicaciones en congresos, informes de investigación, proyectos de final de carrera, etc., y RACO (Revistes Catalanes amb Accés Obert) almacena el texto completo de artículos de revistas científicas, culturales y eruditas catalanas que pueden ser consultados en acceso abierto.

En septiembre de 2006, la Biblioteca Nacional de Catalunya ha puesto en marcha otro ambicioso repositorio en colaboración con CESCA, Patrimoni Digital de Catalunya (PADICAT), que tiene como objetivo capturar, procesar y dar acceso permanente a toda la producción cultural, científica y de carácter general catalana elaborada en formato digital, es decir, archivar la web catalana.

Es importante matizar que estos repositorios son *colectivos*. Integrar en un único catálogo documentos de diversas instituciones es un hecho diferencial e innovador respecto otras iniciativas similares. La participación de diferentes instituciones y universidades facilita la adopción de procedimientos comunes, aumenta la difusión de los contenidos y posibilita que los usuarios realicen consultas globales. Podríamos decir que el total produce mejores resultados que la suma de sus partes.

Un pilar básico sobre el que se han desarrollado los repositorios ha sido el *acceso abierto* a la información, que consiste en poner la producción académica y de investigación en la red de manera libre y gratuita, con el objetivo de crear alternativas al paradigma de pagar para tener acceso a la información que muchas veces se ha elaborado en la propia institución. Un segundo pilar es el uso de *software de código abierto*, que ha permitido la adaptación del *software* a los requisitos concretos de los repositorios.

Desarrollo y plataforma de los repositorios

Para el desarrollo de los cuatro repositorios se ha usado *software* de código abierto^[9]. Para el TDR se adaptó el Electronic Theses and Dissertations (ETD)^[10] de la Universidad Virginia Tech para la gestión de las tesis, y Glimse y WebGlimse^[11] para la indexación y la búsqueda. Para RECERCAT, el DSpace^[12], hecho por el Massachusetts Institute of Technology (MIT) y Hewlett Packard (HP). Y en RACO, se ha usado Open Journal Systems (OJS)^[13] del Public Knowledge Project (PKP).

Como en el caso del TDR, para PADICAT también diferentes *softwares* conviven realizando diferentes funcionalidades. Heritrix^[14] es el encargado de recolectar las páginas web y almacenarlas en archivos comprimidos. Después, el software NutchWAX^[15] realiza el proceso de indexación y también se encarga de realizar las búsquedas que el usuario efectúe, a través de la interfaz de consulta realizada con WERA^[16].

TDR, RECERCAT y RACO son *data providers* de metadatos a través del protocolo de interoperatividad OAI-PMH^[17], característica que hace posible que otros repositorios a nivel internacional puedan recolectar de ellos metadatos, aumentando por consiguiente la visibilidad de su e-información. Además, RECERCAT también es *data service* del mismo protocolo, lo que permite recolectar, no sólo exportar metadatos sino también importarlos.

Los cuatro repositorios están dentro de un clúster Linux de alta disponibilidad con las características de balanceo de carga de las peticiones que les lleguen, y de tolerancia a fallos en caso de desastre en alguno de los nodos que componen la plataforma.

Conclusiones

El TDR, actualmente con 15 universidades, está plenamente consolidado. Dispone de más de 3.500 tesis doctorales en acceso abierto al texto completo, y es un referente en la consulta de contenidos de calidad como lo demuestran las más de 2 millones de consultas recibidas a lo largo del año 2005, un 30% de las cuales venían de países latinoamericanos, con México a la cabeza. En 2006, año de su quinto aniversario, el número de consultas sigue subiendo, con picos de más de 300.000 durante el mes de marzo.

RECERCAT dispone actualmente de 12 instituciones y más de 3.100 documentos de investigación consultables. Todos ellos bajo una licencia Creative Commons^[18] de Reconocimiento-NoComercial-SinObraDerivada. RACO, por su parte, agrupa a 17 instituciones que dan acceso abierto a más de 18.700 artículos agrupados de 108 revistas.

PADICAT sólo había capturado 30 webs en unos 9 GB el día de su puesta en operación pública el 11 de septiembre de 2006. Sin embargo, se tiene la previsión que en 2009 el

proyecto esté plenamente consolidado con 100.000 versiones de web capturadas en lo que podría ocupar 30 TB.

Referencias

- [1] Centre de Supercomputació de Catalunya, <http://www.cesca.es>
- [2] Consorci de Biblioteques Universitàries de Catalunya, <http://www.cbuc.es>
- [3] Tesis Doctorales en Red (TDR), <http://www.tesisenred.net>
- [4] Dipòsit de la Recerca de Catalunya (RECERCAT), <http://www.recercat.net>
- [5] Revistes Catalanes amb Accés Obert (RACO), <http://www.raco.cat>
- [6] Biblioteca Nacional de Catalunya, <http://www.bnc.cat>
- [7] Patrimoni Digital de Catalunya (PADICAT), <http://www.padi.cat>
- [8] Scholarly Publishing and Academic Resources Coalition (SPARC), <http://www.sparceurope.org/Repositories>, <http://www.arl.org/sparc/ir/ir.html>
- [9] "Dipòsits col·lectius d'e-informació". Ricard de la Vega et altri. <http://www.recercat.net/handle/2072/2195>
- [10] Electronic Theses and Dissertations (ETD), <http://etd.vt.edu>
- [11] WebGlimpse, <http://webglimpse.net>
- [12] DSpace, <http://www.dspace.org>
- [13] Open Journal Systems (OJS), <http://pkp.sfu.ca/ojs>
- [14] Heritrix, <http://crawler.archive.org>
- [15] NutchWAX, <http://archive-access.sourceforge.net/projects/nutch>
- [16] WERA, <http://archive-access.sourceforge.net/projects/wera>
- [17] Open Archives Initiative (OAI), <http://www.openarchives.org>
- [18] Creative Commons (CC), <http://creativecommons.org>



Autores

Ricard de la Vega, *rdelavega@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)

Marc Boix, *mboix@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)

Joan Cambras, *jcambbras@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)

Núria Comellas, *ncomellas@cbuc.es* del Consorci de Biblioteques Universitàries de Catalunya

M^a Teresa Novo, *mnovo@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)

Jordi Prats, *jprats@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)

Sandra Reoyo, *sreoyo@cbuc.es* del Consorci de Biblioteques Universitàries de Catalunya

Ramon Ros, *rros@cbuc.es* del Consorci de Biblioteques Universitàries de Catalunya

Natalia Torres, *ntorres@cesca.es* del Centre de Supercomputació de Catalunya (CESCA)