CAT (CURATOR ARCHIVING TOOL): MEJORANDO EL ACCESO A LOS ARCHIVOS WEB

Ciro Llueca, Daniel Cócera
Biblioteca de Catalunya (BC)
Barcelona, España
padicat@bnc.cat

Natalia Torres, Gerard Suades, Ricard de la Vega
Centre de Supercomputació de Catalunya (CESCA)
Barcelona, España
padicat@cesca.cat

ABSTRACT

PADICAT es el archivo web creado en 2005 en Cataluña (España) con el objetivo de capturar, procesar y dar acceso permanente al patrimonio digital de Cataluña. Basa su estrategia de captura en el modelo híbrido (captura masiva del dominio .cat; captura selectiva de los agentes productores de las páginas web catalanas; captura focalizada de acontecimientos públicos). El sistema ofrece su colección en abierto, en Internet. Para hacerlo de manera óptima se ha creído necesario complementar los actuales programas de búsqueda y visualización con una nueva herramienta de software libre, CAT (Curator Archiving Tool), formada por tres módulos orientados a gestionar eficazmente los procesos de catalogación humana; publicar los recursos en directorios y centros de interés temático; y ofrecer a los usuarios información estadística de valor añadido. En el marco del International Internet Preservation Consortium *meeting* (Viena 2010) se presenta a la comunidad internacional los avances en la producción de esta nueva herramienta informática, y la filosofía que ha motivado su diseño.

1. INTRODUCCIÓN

PADICAT (Patrimoni Digital de Catalunya) es el archivo web creado en 2005 en Cataluña (España) por la Biblioteca Nacional de Catalunya (BC) y el Centre de Supercomputació de Catalunya (CESCA), con el objetivo de capturar, procesar y dar acceso permanente al patrimonio digital de Cataluña, entendido este como toda la producción cultural, científica y de carácter general producida en formato digital[1] y publicada en Internet. La misión de PADICAT[2] es archivar la Internet catalana.

La BC forma parte del IIPC (International Internet Preservation Consortium), y como en el resto de proyectos en funcionamiento, el archivo web con sede en Barcelona se basa en la aplicación de una serie de programas informáticos que permiten la captura, el almacenaje, la organización, la preservación y el acceso permanente a una serie de versiones de las páginas web publicadas en Internet por una comunidad determinada[3], en este caso la catalana, en España.

A partir del análisis inicial de los modelos Internet Archive, Kulturarw3 (National Library of Sweden), Pandora (National Library of Australia), y Netarkivet.dk (The Royal Library, State & University Library, Denmark), y de acuerdo con lo que consideramos que es la tendencia generalizada en bibliotecas y archivos nacionales, el modelo de depósito que se persigue en PADICAT es el modelo híbrido, consistente en:

- Compilar masivamente los recursos digitales publicados en abierto en Internet, mediante la captura exhaustiva del dominio .CAT.
- Impulsar el depósito sistemático de la producción web de las entidades catalanas, mediante la identificación y el convenio con las entidades y empresas representativas de la sociedad catalana.
- Promover líneas de investigación mediante la integración temática de los recursos digitales de determinados acontecimientos de la vida pública catalana, com es el caso de las diversas campañas electorales en Internet, el fenómeno de la música en línea, o el mundo de los museos.

[1] Webb, C. *Guidelines for the Preservation of Digital Heritage*. United Nations Educational, Scientific and Cultural Organization, Paris, 2003. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

[2] See "What PADICAT is" at http://www.padicat.cat/en/quees.php

[3] Gomes, D.; Silva, M. J. "Characterizing a National Community Web". *ACM Transactions on Internet Technology*, vol 5, núm 3 (Aug 2005). http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf

PADICAT utiliza en su funcionamiento ordinario Heritrix, Nutchwax, Wera, Wayback Machine, así como el software Web Curator Tool (WCT), desarrollado por la National Library of New Zealand en colaboración con la British Library.

A septiembre de 2010, PADICAT ofrece en su web (http://www.padicat.cat) acceso en abierto a 53.249 capturas de 30.481 sitios web y mantiene convenios de colaboración con 442 administraciones públicas, empresas, universidades, asociaciones profesionales y centros culturales de Cataluña. El acceso ofrecido en abierto a toda la colección, a través de la búsqueda o navegación temática, es justamente uno de los puntos fuertes del archivo web catalán. Conjuntamente con el objetivo de preservar la información publicada en Internet, la visualización de los fondos capturados en Internet ha sido una obsesión: conscientes de las limitaciones legales que afectan a la mayoría de archivos web existentes, PADICAT ha seguido desde su inicio la premisa de permitir el acceso abierto a su colección, defendida por el Internet Archive, a la vez que se ha ido tejiendo artesanalmente una red de entidades y empresas que autorizan, mediante un convenio de colaboración, la difusión de las versiones capturadas en el portal del archivo web.

El objetivo estratégico de garantizar una recuperación óptima y visualización de la información capturada se ha llevado a cabo mediante la asignación de valor añadido a los datos de los registros capturados, asociando nueva información descriptiva a los ítems digitales. En definitiva, catalogando los recursos web que PADICAT ha ido capturando o ha de capturar próximamente. A fecha de hoy se han catalogado 17.700 páginas web.

La información proporcionada por la catalogación manual es la que deberá permitir una mejora sustancial en la identificación y recuperación de los recursos según el título normalizado, el posicionamiento alfabético en el directorio de recursos del repositorio, la pertenencia y la indexación dentro de unas categorías temáticas concretas, la interrelación de los propios recursos dento de la colección de PADICAT, etc. El reto último es la intervención en el scoring de los actuales sistemas de búsqueda (Wera y Wayback) para ponderar los resultados teniendo en cuenta metadatos procedentes de la catalogación.

En definitiva, nuevos elementos de búsqueda que han de ofrecer como resultado final una mejor visibilidad de los recursos que forman parte de la colección, permitiendo al usuario final una interrogación lógica del sistema más allá que conozca o no la URL del recurso deseado. La catalogación de recursos web ha impulsado el diseño y producción de un conjunto de herramientas que llamamos CAT (Curator Archiving Tool).

Es objetivo de este informe presentado en el International International Preservation Consortium *meeting* (celebrado en la 7th International Conference on Preservation of Digital Objects, iPRES2010) dar cuenta a la comunidad internacional de las mejoras planteadas por el equipo de PADICAT a los procesos de recuperación y visualización de la información contenida en los archivos web, así como poner a disposición de los proyectos homólogos los avances en estas líneas de investigación e implementación.

2. CAT (CURATOR ARCHIVING TOOL)

Es evidente que la captura de páginas web publicadas en Internet está suficientemente resuelta por softwares más maduros, como Heritrix. Por contra, los softwares de indexación y los correspondientes módulos de visualización no han evolucionado al mismo ritmo.

Como en el resto de catálogos, la recuperación de la información contenida en un archivo web se realiza bien mediante la búsqueda o bien a través de la navegación en un directorio, habitualmente temático. Solo la parte de búsqueda dispone de programas específicos de amplia difusión en la comunidad, como son Wera y Wayback, a partir de consulta mediante una URL concreta, o de una palabra clave. La navegación por directorios temáticos, planteada en archivos web como Pandora o UK WAC, únicamente puede surgir cuando existen procesos tradicionales de catalogación, come el que se sigue también en PADICAT, con el objetivo último de integrar sus colecciones a los catálogos de la Biblioteca Nacional de Catalunya.

Así pues, las etapas del proceso de gestión de recursos web como la captura o la búsqueda disponen de softwares suficientemente maduros y muy usados por parte de proyectos similares, pero existe un vacío en cuanto a programas para la catalogación y publicación de estos recursos.

Ante este reto, y para garantizar una mejor recuperación y visualización por parte de los públicos externos a los archivos web, el equipo de PADICAT ha llevado a cabo desde 2005 una serie de ensayos e implementaciones que han hecho posible, en la actualidad, presentar una herramienta dirigida a influenciar positivamente en los procesos de búsqueda y recuperación de la información, así como en la presentación de esta información al público que consulta un archivo web.

Se trata de los módulos que forman CAT (Curator Archiving Tool), diseñados con criterios de modularidad, escalabilidad e internacionalización que garantizan su publicación como software libre.

Se ha diseñado y se aborda la producción de un software formado por tres módulos: el básico, de Catalogación (MOCA), que se complementa con más funcionalidades a partir de módulos adicionales: el de Publicación (MOPU), y el de Estadísticas (MOST), que unidos forman la herramienta CAT (Curator Archiving Tool), objeto de la presente comunicación.

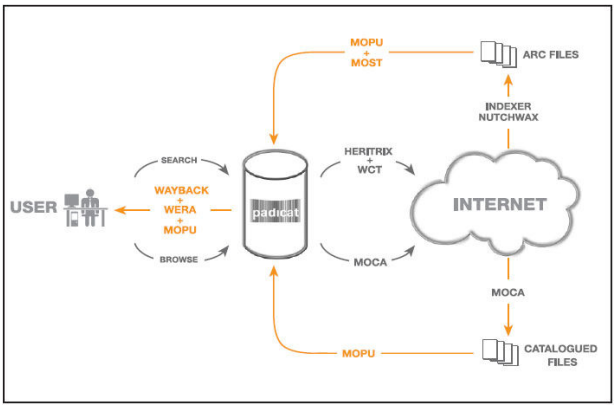


Figura 1. Software utilizado en PADICAT

Veamos a continuación una descripción de los módulos que integran la herramienta CAT (Curator Archiving Tool).

2.1. MOCA (Módulo de Catalogación)

La catalogación de recursos se inició utilizando herramientas ya existentes, esencialmente el pionero Web Curator Tool (WCT), pero con la experiencia se fueron detectando carencias en el software y necesidades particulares que hicieron plantearnos desarrollar un nuevo producto centrado en la catalogación.

Uno de los principales requisitos de MOCA es que permita el registro de un histórico de cambios en los metadatos de los recursos catalogados, puesto que esta es una información que podría ser relevante: el ejemplo más evidente es el cambio de URL asociada al recurso preservado.

La información que se recoge para catalogar los recursos se basa en el modelo de metadatos Dublin Core, pero se puede ampliar fácilmente sin alterar el modelo de datos. En PADICAT los recursos capturados son procesados según el sistema de indexación[4] de la Biblioteca Nacional de Cataluña, en base a una clasificación temática propia del portal PADICAT, y de acuerdo con la procedencia de la instrucción de captura (sea porque pertenece al dominio .cat, a los sitios webs procedentes de convenios de colaboración, a recursos recomendados, o a colecciones monográficas).

Un control limita el acceso a la herramienta a los usuarios identificados mediante usuario y contraseña, y una vez validado el usuario solo tiene acceso a las acciones que puede según su rol asignado. El uso de roles simplifica las páginas de usuaria y permite controlar acuradamente las acciones que puede realizar cada uno de ellos. Se han definido tres roles basándose en los perfiles profesionales del personal implicado en el proceso de catalogación: el gestor (tareas de mantenimiento de la herramienta), el catalogador (introducción y actualización de los recursos catalogados) y el observador (control de los recursos catalogados).

El flujo de trabajo de MOCA obliga a verificar la no existencia de un recurso ya integrado en el catálogo para evitar la duplicación de registros. Una vez superado el primer paso se requiere la entrada de los metadatos mínimos obligatorios y ordenar el almacenaje.

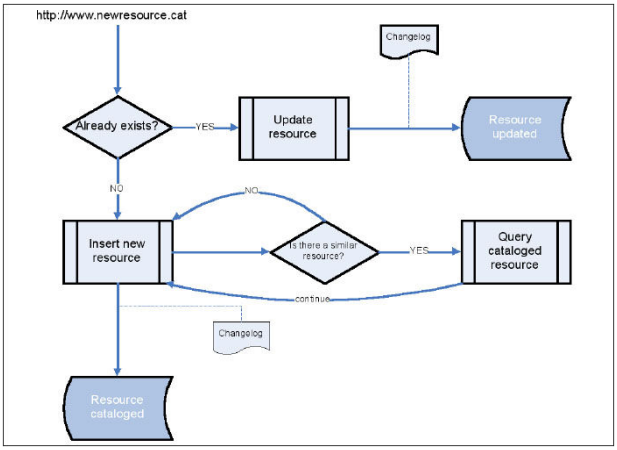
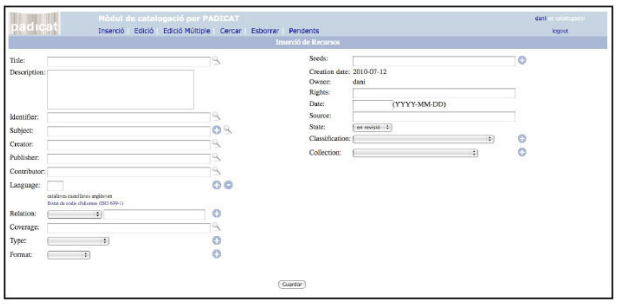


Figura 2. Flujo de trabajo MOCA

A continuación se muestran algunas pantallas de las principales interfaces: la de introducción de metadatos asociadas a un recurso; la de búsqueda que permite hacer búsquedas combinadas utilizando cualquier campo de metadatos; y la de actualización múltiple de registros.



[4] LEMAC, *Llistat d'Encapçalaments de Matèria en Català*, adaptación de las LCSH, *Library of Congress Subject Headings* con aportaciones procedentes del Laval RVM, *Répertoire des Vedettes-Matière*, y el RAMEAU, *Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié*, que la BC mantiene.

Figura 3. Interfaz de inserción de recursos MOCA v1.0

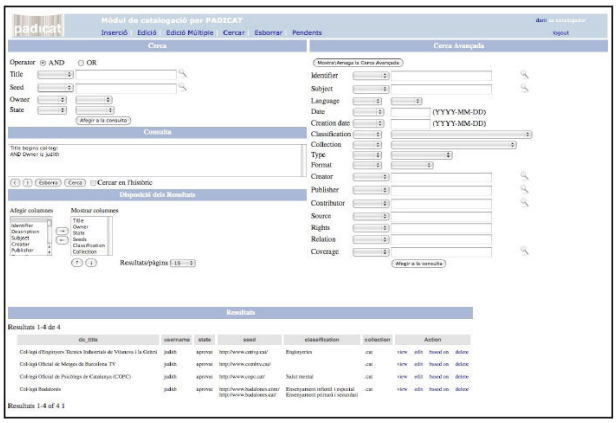


Figura 4. Interfaz para la búsqueda de recursos MOCA v1.0

Una vez desarrollada esta herramienta se han creado unos scripts que han permitido migrar la información existente actualmente de Web Curator Tool hacia el nuevo modelo de datos de MOCA.

La fecha límite prevista para la publicación en abierto de la versión 1.0 de MOCA es a finales de marzo de 2011.

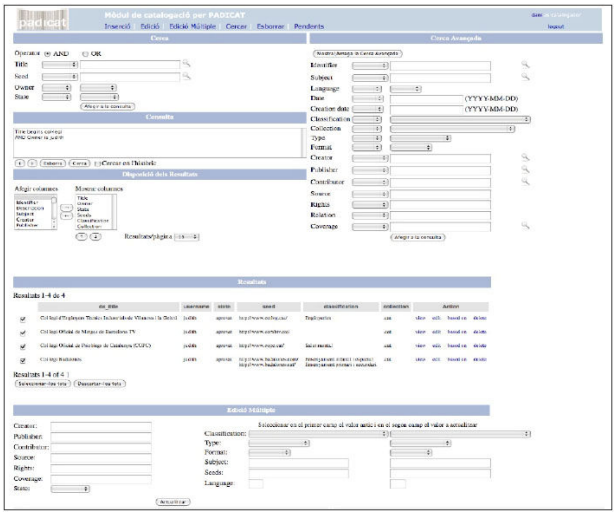


Figura 5. Interfaz de actualización múltiple MOCA v1.0

2.2. MOPU (Módulo de Publicación)

MOPU permitirá generar automáticamente el directorio temático donde estarán disponibles aquellos recursos catalogados previamente con MOCA que hayan sido seleccionados para ser publicados.

Opcionalmente cada uno de estos recursos podrá tener asociada una ficha web, donde se mostrará información descriptiva del recurso procedente de los metadatos de catalogación que se consideren relevantes para la visualización del recurso capturado. Los metadatos que se incluyen en esta ficha son seleccionados de forma individualizada y personalizados para cada uno de los recursos. Es imprescindible disponer de MOCA para poder utilizar MOPU.

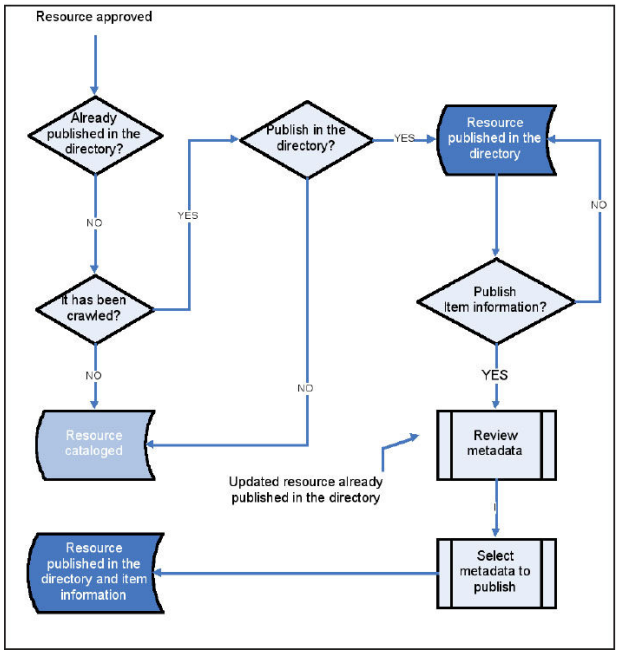


Figura 6. Flujo de trabajo MOPU

Hace falta definir un nuevo rol para MOPU, el validador. Este nuevo rol de usuario será el encargado de gestionar todo el proceso de publicación. Comprobará los datos introducidos por el catalogador y dará el beneplácito a la publicación del recurso en el directorio temático y si es oportuno, qué metadatos se visualizarán en la ficha web. En el caso que detecte alguna carencia en los metadatos del recurso, el validador marcará el recurso para que lo revise el catalogador y a través de un campo de notas asociado al recurso comunicará el motivo de los cambios a realizar.

Paralelamente se generará de forma automática un histórico de cambios que permitirá hacer un seguimiento sobre la actividad de publicación y los cambios en la ficha web del recurso.

A continuación se muestran algunas de las pantallas de las interfaces según el piloto desarrollado: la interfaz de gestión del directorio temático según las categorías; y la interfaz para la selección de los metadatos a incluir en la ficha web de un recurso.

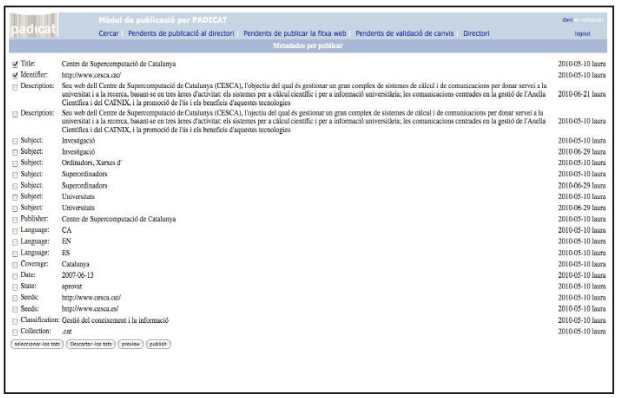


Figura 7. Interfície del prototip MOPU

Figura 8. Interfaz del prototipo MOPU

2.3. MOST (Módulo de Estadísticas)

Este módulo permitirá generar estadísticas sobre la información incluida en el repositorio a partir de los informes generados por el software de captura, siendo ejemplos el nombre y la tipología de ficheros que forman cada una de las capturas preservadas, volumen y periodicidad de cada captura, etc. Este es un servicio de valor añadido ya que desde PADICAT se pueden consultar en abierto todos los recursos capturados.

La integración de este módulo con los anteriores permitirá incluir estadísticas por recurso en la ficha web de presentación de los recursos preservados.

2.4. Hoja de ruta

Tan pronto como se disponga de una versión estable y bastante madura de los módulos que forman CAT (Curator Archiving Tool), con el feedback de su utilización, se ofrecerá a la comunidad como software libre. En el caso de MOCA está previsto que sea antes del 31 de marzo de 2011.

3. CONCLUSIONES

La captura de páginas web publicadas en Internet está suficientemente resuelta por softwares ya maduros, como Heritrix. Por contra, los softwares de indexación y los correspondientes módulos de visualización no han evolucionado al mismo ritmo. Para archivos web como PADICAT, que ofrecen en Internet su colección en abierto, es imprescindible mejorar los procesos de búsqueda y visualización.

La concepción de la herramienta CAT (Curator Archiving Tool) permite mejorar la búsqueda y visualización de los recursos preservados en archivos web, gracias a sus aportaciones en:

- La incidencia humana en la descripción e indexación mediante el módulo de Catalogación.
- La generación automatizada de directorios y centros temáticos de interés mediante el módulo de Publicación.
- El incremento en la oferta de datos estadísticos mediante el módulo de Estadísticas.

Para los gestores del archivo web, la interacción de los tres módulos permitirá beneficios en ahorro de tiempo derivado de la automatización de procesos actualmente manuales:

- Integrar automáticamente los recursos preservados en listas alfabéticas, en directorios temáticos o en centros de interés, que conforman el acceso a la colección mediante la navegación.
- Editar en grupo recursos catalogados que compartan las mismas características.
- Mejorar los flujos de trabajo entre los diferentes roles que operan en la catalogación y publicación de recursos en el portal de acceso al archivo web.
- Adaptar PADICAT a la estrategia de la Biblioteca Nacional de Catalunya, con el inicio de la integración de los recursos preservados en los catálogos y sistemas de búsqueda ordinarios de la biblioteca.

Para los usuarios del archivo web, la creación y puesta en funcionamiento de la herramienta CAT (Curator Archiving Tool) permitirá una mejora en la búsqueda y visualización de los recursos preservados:

- Buscar por navegación temática más eficazmente y exhaustivamente, complementando los sistemas ya existentes de búsqueda por palabra clave o bien por URL.
- Buscar por palabra clave de manera más pertinente, una vez se intervenga en el scoring de los sistemas

de búsqueda, con datos procedentes de la catalogación.
- Buscar integradamente, una vez se integre la colección de PADICAT en los catálogos y sistemas de búsqueda ordinarios de la Biblioteca de Catalunya.
- Acceder a información de valor añadido para cada recurso preservado, sea procedente de la descripción humana de los recursos, sea procedente de los datos estadísticos que crea el sistema de funcionamiento.

Para la comunidad internacional, la publicación en abierto de los módulos de la herramienta CAT (Curator Archiving Tool) permite:

- Contribuir a la misión de adquirir, preservar y hacer accesible la información de Internet para futuras generaciones, por todo el mundo, promoviendo el intercambio global y las relaciones internacionales.
- Potenciar la intervención humana en los procesos de descripción y publicación de los archivos web para mejorar el acceso y visualización de estos sistemas.
- Conseguir reconocimiento público y político a la utilidad de proyectos de preservación digital de Internet, como son los archivos web protagonistas de la cita anual en el IIPC *meeting*.