12as Jornadas Españolas de Documentación - Fesabid 2011

Eje temático: Una profesión que innova

Título:

PADICAT, el archivo de Internet

Autores:

Ciro Llueca, Daniel Cócera, Natalia Torres, Gerard Suades, Ricard de la Vega

Datos de contacto:

PADICAT (Patrimonio Digital de Cataluña)

Biblioteca de Catalunya

Hospital, 56 - 08001 Barcelona

cllueca@bnc.cat

Resumen:

PADICAT es el archivo web creado en 2005 por la Biblioteca de Catalunya con el objetivo de capturar, procesar y dar acceso permanente al patrimonio digital de Cataluña en Internet. Basa su estrategia de captura en el modelo híbrido (captura masiva del dominio .CAT; captura selectiva por convenio de los agentes productores de las páginas web catalanas; captura focalizada de acontecimientos públicos). El repositorio ofrece su colección en abierto, en Internet. Tras 5 años de experiencia, se describe el sistema de funcionamiento y los retos de futuro.

Abstract:

PADICAT is the web archive created in 2005 by the Biblioteca de Catalunya (BC), with the aim of collecting, processing and providing permanent access to the digital heritage of Catalonia (Spain) on the Internet. Its harvesting strategy is based on the hybrid model (massive harvesting of .CAT top level domain; selective compilation by agreement of the web site of organizations; focused harvesting of public events). The repository provides open access to the whole collection, on the Internet. After 5 years of experience, we describe the operating system and future challenges.

Palabras clave:

Repositorios digitales; Bibliotecas digitales; Archivos web; Preservación digital.

Keywords:

Digital repositories, Digital libraries, Web archives; Digital preservation.

Texto completo:

1. Antecedentes

Desde la década de los noventa, y a partir de la aparición progresiva de servidores y páginas web en Internet, las administraciones públicas de diversos países han llevado a cabo estrategias para garantizar el acceso y preservación de los contenidos publicados en Internet: el patrimonio digital[1]. El reto no es menor: además de la inexistencia generalizada de un texto legal actualizado[2] que dé cobertura a estos procesos documentales, no existen sistemas informáticos que ejecuten impecablemente las operaciones de compilación, proceso y difusión de las páginas web en un entorno, Internet, que es dinámico por naturaleza.

Pese a estas dificultades, varios países están realizando acciones de preservación de la producción digital más obvia, las páginas web, mediante la creación de repositorios digitales llamados comúnmente “archivos web”. Existen diversos archivos web en funcionamiento, así como extensa bibliografía que los ha detallado y analizado[3]. Los más conocidos son también los que dieron los primeros pasos en 1996: el sueco *Kulturarw3* y el australiano *Pandora*; así como un repositorio de alcance internacional, el gigante *Internet Archive*. Quince años más tarde podemos contar hasta 36 proyectos en diversas fases de implementación, siendo acciones consolidadas un tercio de esa cifra[4]. El análisis de estas experiencias muestra dos modelos básicos de políticas con una tendencia generalizada hacia un modelo híbrido. El primero es el modelo integral o exhaustivo (mayoritario, y característico especialmente de los países escandinavos), que persigue la integración automática de la Web a partir de determinados criterios infraestructurales (según el dominio de las páginas web, según la ubicación del servidor, lingüísticos, etc.). El segundo modelo es el selectivo (asimilado por Australia, el Reino Unido o Japón, entre otros países), dirigido a compilar la Web en base a una política selectiva (sobre un espacio geográfico determinado, un tema de interés nacional, etc.). Estos dos modelos han dado paso, en lo que es ya una tendencia generalizada, a modelos híbridos (cuyo caso más evidente es el de Dinamarca) que complementan la captura periódica de la Web con acciones selectivas, ampliando esa cobertura a determinados acontecimientos de interés social (elecciones, competiciones deportivas, etc.). Lamentablemente, el número de depósitos que permiten acceder libremente a sus colecciones o a su fondo es muy limitado, bien porque se quieren evitar conflictos con la vulneración de los derechos de propiedad intelectual de los recursos capturados sin autorización expresa[5], bien porque las interfaces de recuperación de la información depositada no se han desarrollado eficazmente.

En la mayoría de los casos han sido impulsores de estos proyectos las bibliotecas y archivos nacionales. En el caso español la Biblioteca de Catalunya (BC) puso en funcionamiento en 2005 el repositorio PADICAT (Patrimonio Digital de Cataluña)[6], dedicado al archivo sistemático de la Internet *catalana*[7]. En 2007 el Gobierno Vasco y EJIE (Sociedad Informática del Gobierno Vasco) crearon Ondarenet[8], Archivo Electrónico del Patrimonio Digital Vasco. Desde 2009, la Biblioteca Nacional de España encarga capturas periódicas del dominio .es a Internet Archive, con sede en Estados Unidos.

2. Desarrollo de PADICAT

En junio de 2005, coincidiendo prácticamente con su centenario, la BC impulsó acciones para evolucionar hacia un modelo de biblioteca abierta, fiable y orientada al usuario. En el marco de su misión[9], una de las líneas estratégicas ha sido el impulso de proyectos digitales[10] de carácter eminentemente cooperativo para contribuir a la preservación del patrimonio catalán y aumentar la presencia de contenidos catalanes en Internet. Algunos de esos proyectos son *ARCA* (Archivo de Revistas Catalanas Antiguas), *Memòria Digital de Catalunya*, *RACO* (Revistas Catalanas de Acceso Abierto), *CLACA* (Clásicos Catalanes), el proyecto *Google Books* de digitalización del fondo libre de copyright de cinco bibliotecas patrimonialistas catalanas[11] o el propio PADICAT, un depósito digital que cuenta con la colaboración del CESCA (Centre de Supercomputació de Catalunya) y de la Generalitat de Catalunya y con un presupuesto de 800.000 euros (2005-2011). Basa su política de colección en el modelo híbrido, orientado a:

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

Por lo que se refiere a la arquitectura del sistema, tras la fase de análisis y testeo de programas se determinó que se utilizaría el programa informático Heritrix[12], usado por la mayoría de proyectos para la captura de recursos digitales. Este programa es el encargado de recolectar las páginas web tal como las ve el usuario que navega por Internet, y almacenarlas en archivos comprimidos en formato ARC[13]. Después, NutchWax[14] y Hadoop[15] realizan un proceso de indexación de la información recolectada que permitirá, posteriormente, utilizar estos índices para localizar recursos dentro de la colección. Existen dos interfaces para realizar las consultas al conjunto de recursos

capturados y acceder a su visualización: Wera[16], que permite la búsqueda por palabras clave a través de los índices generados por NutchWax; y Wayback[17], que permite la consulta directa por URL. Finalmente, el programa Web Curator Tool[18], desarrollado por la National Library of New Zealand y la British Library, se ha aprovechado como sistema de gestión documental que permite la asignación de metadatos a una parte significativa de la colección, lo que garantiza la posibilidad futura de integrar la colección en otros catálogos de la BC u otras instituciones (aunque para ello, como veremos, será necesaria una evolución considerable de las actuales prestaciones informáticas). El personal del CESCA, socio tecnológico del proyecto, ha desarrollado y compartido con la comunidad diversas aplicaciones *ad hoc*, como los módulos del CAT (Curator Archiving Tool)[19] diseñados para mejorar el acceso y recuperación de los recursos digitales depositados en PADICAT. Todo el *software* utilizado es de código abierto y gratuito.

Por lo que afecta a la preservación digital, somos conscientes de la problemática de las estrategias más habituales de preservación[20], como la migración periódica o *refresh* de los datos (migración a nuevas versiones de los mismos programas o lenguajes, o a nuevos programas capaces de leer los anteriores), la emulación (el uso de *software*, especificaciones, etc., utilizado en el momento de la creación), la recreación (simulación por ingeniería inversa u otros métodos). En todo caso, las previsiones sobre el tipo de archivos que el repositorio debe gestionar, basadas en la actual composición del fondo de la colección, revelan que la mayor parte de los archivos corresponden a formatos estándares, que pueden simplificar la tarea preservadora al menos en las macrocifras. Así, sobre una muestra de 250 millones de ficheros, aproximadamente el 95% de ellos corresponden a formatos estándares: texto/html (82%), imagen jpeg o gif o png (12%), o pdf (1%). El resto de formatos (audio, vídeo, flash, etc.) tiene presencial residual.

En cuanto al *hardware* utilizado, PADICAT tiene a su disposición cinco nodos HP ProLiant DL360 G4p encargados de las tareas de recolección e indexación de webs. De la búsqueda y visualización de resultados en la interfaz web se encarga un *clúster* Linux de alta disponibilidad con características de balanceo de carga de peticiones y de tolerancia a fallos en caso de desastre en los nodos que componen la plataforma. Una cabina NetApp FAS3170 presenta un espacio de disco via NFS a estos nodos. El sistema se completa con un robot donde se guardan copias de seguridad de los datos en cinta. En el momento de presentación de la presente experiencia profesional un depósito de preservación digital a largo plazo desarrollado por la BC se halla en fase piloto.

En septiembre de 2006 se inauguró la web de PADICAT (www.padicat.cat) en una versión trilingüe que hoy, corregida y aumentada, se mantiene. Desde el primer día, como filosofía del repositorio, se ha dado acceso abierto a toda la colección disponible en el depósito. Primero, con un motor de

búsqueda a texto completo. En una segunda fase, con la creación de centros de interés monográficos. Finalmente, se han completado las opciones anteriores con la opción de búsqueda a través de URL y, sobretodo, de un directorio temático dirigido al público que prefiere la navegación como fórmula de visita del fondo que forma la colección de PADICAT.

3. Conclusiones y perspectivas

El archivo de Internet es posible. PADICAT, tras 5 años de existencia, contiene 39.587 capturas de 118.039 páginas web y está formado por 249 millones de ficheros, con un tamaño de 7,5 TB[21]. Se han firmado acuerdos de colaboración con 450 instituciones y empresas. Y se han creado 7 monográficos dedicados a las campañas electorales en Internet, música folk-rock, y museos catalanes. Consolidada la infraestructura técnica, la previsión de crecimiento anual para 2011 se establece en 40.000 capturas, 50 nuevos convenios, y 2 monográficos. Lo más importante: se está realizando un trabajo sistemático de compilación, procesamiento y difusión del patrimonio digital de Cataluña en Internet.

El futuro, después de una etapa que podemos considerar de nacimiento, pasa por consolidar su capacidad de crecimiento, mejorar sus procesos de trabajo y optimizar sus recursos:

- Consolidando la infraestructura necesaria del repositorio, para dar respuesta tecnológica al reto del crecimiento exponencial que perseguimos.
- Definiendo las estrategias de preservación digital, como uno de los aspectos clave de transferencia de conocimiento a la sociedad.
- Potenciando las recopilaciones monográficas sobre la vida pública catalana en Internet, y aprovecharlas para trazar alianzas estratégicas con colectivos de expertos según disciplinas.
- Creando la hemeroteca digital catalana en Internet, por medio de la captura sistemática, diaria, de las publicaciones seriadas en Internet.
- Cooperando con otros archivos web e instituciones de la memoria, para dar respuesta eficiente a los retos de preservación digital y acceso a los recursos depositados.

En la presentación de los objetivos de PADICAT, en 2005[22], se planteaban los potenciales beneficios de un proyecto que se encontraba en un estadio preliminar. Cinco años más tarde, los beneficios son plenamente vigentes desde el momento en que han llegado a ser factores críticos de éxito en la estrategia de la BC: para la comunidad bibliotecaria, los beneficios se centran en la integración de los documentos nacidos digitales en la bibliografía nacional, y su difusión como fuente de información de los documentos que representan el futuro; la creación de amplios escenarios de cooperación con las instituciones de la memoria: bibliotecas, archivos y museos, así como universidades y centros de investigación. Para las instituciones, empresas, administraciones

y particulares que producen páginas web en Cataluña, preservación de la propia producción y garantía de acceso, con los condicionantes que rige la ley, a los contenidos y diseños que, de otra forma, desaparecerían. Y por último, para la ciudadanía, y como se pretende en las Directrices de la Unesco, acceso abierto y permanente a los recursos que forman el patrimonio digital.

4. Referencias

[1] *Directrices para la preservación del patrimonio digital*. Canberra: Unesco, 2003. [Consulta: 01/12/2010] http://unesdoc.unesco.org/images/0013/001300/130071s.pdf

[2] Aunque parece inminente una actualización, el texto legal español vigente data de 1971. Es ejemplo paradigmático de buena práctica la ley danesa del depósito legal, de 2004. Véase una traducción al inglés en: Act on legal deposit of published material: translation of Act N. 1439 of 22 December 2004: unauthorized version. [Consulta: 01/12/2010] http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html, así como un exhaustivo informe comparativo: Georgia, A. *Digital legal deposit in the EU member states: an overview of regulatory and implementation status: background report in connection on digital libraries*. Frankfurt: Foundation Conference of European National Librarians, 2006. [Consulta: 01/12/2010] http://web3.nlib.ee/cenl/docs/Digital%20Legal%20Deposit%20in%20the%20EU%20member%20states.pdf

[3] Para un panorámica global sobre estos proyectos véase: Llueca, C. "Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales". *BiD: textos universitaris de biblioteconomia i documentació*, 2005, diciembre, n. 15. [Consulta: 01/12/2010] http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec2.htm

[4] La mayor parte de los proyectos, incluyendo PADICAT, se hallan representados en el International Internet Preservation Consortium (IIPC): http://netpreserve.org/

[5] Numerosos proyectos han optado por no mostrar en acceso abierto sus colecciones, por el temor a vulnerar las leyes de propiedad intelectual. Internet Archive, y los archivos web portugués, japonés, británico, australiano, israelí, vasco y catalán, entre otros, muestran en abierto sus colecciones, bien por haber llegado a acuerdos expresos con los productores de las páginas web (y propietarios de los derechos), bien por dar cumplimiento a la filosofía de acceso abierto que ha reflejado Josep Vives en "Aspectos de propiedad intelectual en la creación y gestión de repositorios institucionales". *El profesional de la información*, 2005, julio-agosto, v. 14, nº 4. http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/267.pdf, en el sentido de que "disponemos de buenos y suficientes argumentos para convencer a nuestros depositantes de la bondad de los repositorios, sin entrar en debates estériles sobre la legalidad o no de preservar la producción digital. A título de ejemplo, ni en Suiza ni en los Países Bajos existen siquiera leyes de depósito legal".

[6] El portal PADICAT, http://www.padicat.cat está operativo desde 2006.

[7] Interesante reflexión sobre comunidades nacionales en Internet en: Gomes, D.; Silva, M. J. "Characterizing a National Community Web". *ACM Transactions on Internet Technology*, vol 5, núm 3 (Aug 2005). [Consulta: 01/12/2010] http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf

[8] El portal Ondarenet, http://www.ondarenet.kultura.ejgv.euskadi.net está operativo desde 2007.

[9] Según consta en las leyes catalanas de bibliotecas de 1981 (DOGC 123, 29/04/1981) y 1993 (DOGC 1727, 29/03/1993), la Biblioteca de Catalunya tiene por misión recopilar, conservar y difundir la producción bibliográfica catalana y la relacionada con el ámbito lingüístico catalán, y velar por la conservación y la difusión del patrimonio bibliográfico. Entendemos que este patrimonio bibliográfico incluye también la producción bibliográfica digital, la publicada en Internet.

[10] La estrategia y los proyectos, a excepción del Google Books que fue un acuerdo posterior, fueron presentados en: Lamarca, D.; Serra, E. "L'estratègia de la Biblioteca de Catalunya en projectes digitals". *Ítem*, 2005, setembre-desembre, n. 41, pp. 41-43. [Consulta: 01/12/2010] http://www.raco.cat/index.php/Item/article/view/40866/68116

[11] Biblioteca de Catalunya, Biblioteca del Ateneu Barcelonès, Biblioteca Pública Episcopal del Seminari de Barcelona; Biblioteca del Centre Excursionista de Catalunya, y Biblioteca de la Abadia de Montserrat.

[12] Heritrix: http://crawler.archive.org. Veáse también: Mohr, G. [et al.] "An introduction to Heritrix: an open source archival quality web crawler". *International Web Archiving Workshop* (2004). [Consulta: 01/12/2010] http://www.iwaw.net/04/Mohr.pdf

[13] Arc File Format: http://en.wikipedia.org/wiki/ARC_(file_format)

[14] NutchWax: http://archive-access.sourceforge.net/projects/nutch/

[15] Hadoop: http://hadoop.apache.org/core/

[16] Wera: http://archive-access.sourceforge.net/projects/wera/

[17] Wayback: http://www.archive.org/web/web.php

[18] Web Curator Tool: http://webcurator.sourceforge.net/

[19] Más información en el informe técnico: Llueca, C.; Cócera, D.; Torres, N.; Suades, G.; De la Vega, R. "CAT (Curator Archiving Tool): mejorando el acceso a los archivos web". *International Internet Preservation Consortium meeting* (Viena 2010). [Consulta: 01/12/2010] http://www.recercat.net/bitstream/2072/85525/6

[20] Ayre, C.; Muir, A. "The right to preserve: the rights issues of digital preservation". *D-Lib magazine*, v. 10, n. 3 (march 2004). [Consulta: 01/12/2010] http://www.dlib.org/dlib/march04/ayre/03ayre.html

[21] Datos a 23/02/2011. Se hallan actualizados en: http://www.padicat.cat/es/estadistiques.php. Cada página web tiene un régimen específico de capturas, siendo la norma habitual una captura semestral de cada recurso digital integrado en el repositorio. Para los recursos procedentes de monográficos (campañas electorales, etc.) se refuerza el calendario de capturas, según tipología de página web, hasta llegar a la captura diaria de determinados recursos digitales.

[22] Información completa en: Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: Biblioteca de Catalunya, 2005. [Consulta: 01/12/2010] http://www.recercat.net/handle/2072/1757 y sintética en: Llueca, C.

"Archivando la Web, el proyecto Padicat (Patrimonio Digital de Cataluña)". *El profesional de la información*, v. 15, núm. 6 (2006), p. 473-478. [Consulta: 01/12/2010] http://eprints.rclis.org/archive/00007767/01/epi_padicat.pdf