Archivando la Web catalana: iniciativas cooperativas de preservación digital en Catalunya[1]

Eugènia Serra Aranda
Julio 2006

1. El Patrimonio digital: qué es y porqué se debe conservar

Desde la aparición de las Tecnologías de la Información y de la Comunicación, y muy especialmente de la Web, los documentos, expresiones culturales, artísticas y la información en general se generan, diseñan y publican en Internet.

Internet ha cambiado los hábitos sociales, de aprendizaje y trabajo, ha proporcionado nuevas formas de comunicar y generar conocimiento y documentación, y ha generado un contexto y unos productos que, lejos de desaparecer, tienden a crecer exponencialmente y desarrollarse rápidamente.

Si bien en una primera etapa se trataba en muchos casos de exponer o divulgar por Internet contenidos que ya existían en formato analógico (por ejemplo en papel) cada vez son más los contenidos que nacen digitales (*born digital*) y que no tienen un equivalente en el mundo analógico.

La facilidad de crear editar y publicar en Internet, el bajo coste de la edición digital y la capacidad de llegar a cualquiera en cualquier punto del planeta y en cualquier momento, han propiciado el éxito de Internet y de la Web. La tan usada frase de "si no estás en Internet es que no existes" es casi una realidad incuestionable, o como mínimo lo es si nos ceñimos a los países desarrollados que se encuentran en lo que venimos a llamar Sociedad de la Información y del Conocimiento.

Desde hace siglos que los países se preocupan en garantizar la pervivencia de su patrimonio a través de las instituciones de la memoria, o sea de archivos, museos y bibliotecas. Actualmente estas mismas instituciones se encuentran ante la necesidad y el reto de garantizar el acceso permanente a su patrimonio digital.

El patrimonio digital consiste de acuerdo con la definición de la UNESCO[2] en "recursos únicos que son fruto del saber o la expresión de los seres humanos. Comprende recursos de carácter cultural, educativo, científico o administrativo e información técnica, jurídica, médica y de otras clases, que se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente. Los productos "de origen digital" no existen en otro formato que el electrónico".

Existe pues un volumen creciente de contenidos e informaciones nacidos digitales que conjuntamente con el patrimonio bibliográfico analógico constituyen el patrimonio de una sociedad, país o cultura.

---

[1] Ponencia presentada en el curso "LA RECUPERACIÓN DE LA MEMORIA, MUCHAS MÁS OPORTUNIDADES QUE REALIDADES: EL TRABAJO COOPERATIVO DE ARCHIVOS, BIBLIOTECAS Y MUSEOS". Universidad del País Vasco. 23 a 25 de Agost de 2006.

[2] UNESCO. *Carta sobre la preservación del patrimonio digital.* 2003 http://portal.unesco.org/es/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

2. Qué se está haciendo al respecto: una aproximación a la situación a escala internacional

Aproximadamente hacia 1996 surgieron las primeras iniciativas. Las bibliotecas nacionales de algunos países alertadas por el crecimiento exponencial de la Web, se plantearon la necesidad de identificar los recursos electrónicos que se estaban creando y publicando en la Web y establecer los mecanismos para su preservación. Las bibliotecas nacionales tienen entre sus funciones específicas la identificación, recogida, gestión, conservación y difusión del patrimonio bibliográfico.

El patrimonio bibliográfico hasta hace unos cuantos años consistía en libros, revistas, folletos mapas, grabados, dibujos, grabaciones sonoras y videos… y se habían habilitado los procedimientos en muchos casos de carácter legal para garantizar precisamente la recopilación del patrimonio. Por ejemplo en el caso de España existe la obligación de que los impresores o editores depositen de cada documento que publican un número determinado de ejemplares en las respectivas bibliotecas nacionales.

En muchos países la ley de Depósito Legal fue redactada y aprobada mucho antes de la aparición de la Web, por lo cual el patrimonio digital no queda contemplado por la legislación. La ley española que regula el depósito legal es del año 1958[3] y el reglamento posterior de 1974, son pues leyes que en el contexto actual resultan obsoletas.

Suecia y Australia fueron los primeros países en abordar la cuestión del patrimonio digital, y lo hicieron con enfoques y objetivos diferentes, que se han constituido como modelos o referentes para otros países[4].

La Biblioteca Nacional de Suecia se planteó desde el inicio la necesidad de realizar una captura exhaustiva de toda la producción digital que encontraba en la Web sueca. Pese al volumen ingente de información y datos que suponía sostenían que era necesaria la exhaustividad ya que difícilmente se podía seleccionar qué era o no relevante no sólo en la actualidad sino especialmente para las generaciones futuras.

Con esta perspectiva pusieron en marcha el proyecto *Kulturarw3*[5], que en esencia pretendía capturar y almacenar todas las Webs y recursos electrónicos que se hallaban

[3] *Reglamento del Servicio de Depósito Legal* (Decreto de 23 de diciembre de 1957 (BOE 20/1/58)). Orden Ministerial de 22 de noviembre de 1973 que rectifica algunos artículos del anterior. Orden Ministerial de 20 de febrero de 1973, por la que se modifica el *Reglamento del Instituto Bibliográfico Hispánico* (BOE, 3/3/1973). Ordre Ministerial de 30 d'Octubre de 1971, por la que se aprueba el *Reglamento del Instituto Bibliográfico Hispánico* del que el capítulo II regula el Depósito Legal (BOE, 18/11/1971).
Desde hace ya años que el sector bibliotecario reivindica la necesidad de aprobar una nueva ley del Depósito legal que responda a las características y contexto del mercado editorial y de la creación de contenidos digitales. La ley debería ajustarse también al cambio de roles y responsabilidades fruto de la edición digital en Internet. La ley vigente del Depósito Legal española actua sobre los impresores en vez de los editores. Actualmente se debería tener en cuenta en el ámbito de Internet a los creadores de contenidos digitales o a los responsables intelectuales, ya que en muchos casos actúan directamente como editores.

[4] Se puede consultar una panorámica sobre los diferentes proyectos de conservación del patrimonio digital en el artículo:. Llueca, C. (2005). "Webs sempre accessibles : les biblioteques nacionals i els dipòsits digitals nacionals". *BiD: textos universitaris de biblioteconomia i documentació*, núm. 15 (des 2005). (http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec1.htm)

bajo el dominio .se. Desarrollaron un motor (también conocido como "crawler"o "spider") que rastreaba por la red y cuando identificaba un recurso digital que estaba bajo el dominio sueco lo capturaba, indexaba y guardaba una copia de todos los ficheros que comprendía. Lógicamente este rastreo y captura requería tiempo, con lo que el sistema que habían diseñado era capaz de hacer estas capturas 2 veces al año. Teniendo en cuenta lo cambiantes que son las Webs (a veces a diario), si bien no se recogen cada una de las versiones de una web, si que proporcionaba una foto fija completa de la Web sueca.

Paralelamente surgía en Australia el proyecto *Pandora*[6], promovido por la Biblioteca Nacional del país. Pandora partía de unas premisas opuestas a Suecia, reconociendo la dificultad o casi imposibilidad de hacer una recopilación exhaustiva del patrimonio digital, optaron por un modelo selectivo. Este modelo garantiza una gran coherencia en la colección de recursos, ya que son evaluados y seleccionados previamente pero proporciona una visión sesgada o parcial de la Web australiana.

Con el paso de los años, los sucesivos proyectos impulsados desde otros países, si bien inicialmente se han situado en una u otra "escuela", han ido tendiendo –así como también las mismas Australia y Suecia- hacia un modelo híbrido. Hay que tener en cuenta también que la tecnología ha ido avanzando y ofreciendo mejores prestaciones tanto para la captura como para la organización y acceso a los recursos compilados.

El ejemplo más característico del modelo híbrido que se está imponiendo y que ha sido la referencia para el proyecto de Cataluña, es el caso de Dinamarca[7]. El repositorio *Netarkivet.dk* realiza 4 capturas anuales del dominio .dk, captura más frecuente (a veces diaria) de alrededor de 80 dominios de actualización de contenidos elevada -de acuerdo con los editores responsables de su producción- y recoge puntualmente recursos relacionados con hechos o acontecimientos de la historia, cultura y sociedad danesa.

Desde el punto de vista legislativo Dinamarca cuenta con una ley de depósito legal de 1997 que ya contemplaba el depósito de recursos Web, pero el paso definitivo se produjo en diciembre de 2004 cuando el parlamento danés aprobó una revisión de dicha ley por la que se autorizaba a sus bibliotecas nacionales[8] a capturar y preservar la Web danesa.

En la actualidad son ya más de 16 países que en los últimos diez años han puesto en marcha proyectos para archivar y preservar su patrimonio digital[9]

Cierto es que la preservación del patrimonio digital Web es un tema que preocupa al colectivo bibliotecario y cada vez más está presente en foros, seminarios o congresos de

---

[5] *Kulturarw*^{3}*- Long time preservation of electronic documents.* National Library of Sweden (http://www.kb.se/kw3/ENG/)

[6] *PANDORA. Australia's Web Archive*. National Library of Australia. (http://pandora.nla.gov.au/index.html)

[7] *Netarkivet.dk.* The State and University Library; the Royal Library. (http://netarchive.dk/index-en.php)

[8] Si bien es más frecuente que solamente haya una biblioteca nacional en cada estado, en Dinamarca, como en algunos otros países europeos (Italia o Reino Unido, por ejemplo), tanto la State and University Library como la Royal Library tienen rango de biblioteca nacional.

[9] La misma National Library of Australia mantiene la Web *PADI: Preserving Acces to Digital information* (http://www.nla.gov.au/padi/index.htm), que ofrece información actualizada sobre los proyectos, tecnologías, publicaciones y novedades en general en el ámbito de la preservación digital.

nuestro país e internacionales, pero en el caso de España queda aún mucho camino por recorrer.

Recientemente, el pasado marzo, se celebraron las *Jornadas sobre Preservación del Patrimonio Digital*[10] en Madrid impulsadas por la Subdirección General de Cooperación Bibliotecaria del Ministerio de Cultura precisamente orientadas a sensibilizar a los responsables máximos de la cultura y bibliotecas del Estado y de las Comunidades Autónomas sobre la necesidad de poner en marcha en España una iniciativa similar a las ya existentes en el mundo de forma cooperativa o federada. Con este objetivo estuvieron presentes representantes de algunos de los proyectos más importantes de Europa, y también tuve la oportunidad de presentar el proyecto PADICAT (Patrimoni Digital de Catalunya), el primer proyecto de preservación de la Web en España, que ha impulsado la Biblioteca de Catalunya.

3. El proyecto PADICAT (Patrimoni Digital de Catalunya)

Antes de empezar mi exposición sobre el proyecto permítanme que les presente brevemente la Biblioteca de Catalunya, la institución que impulsa y lidera PADICAT, en la cual tengo el privilegio de trabajar.

3.1 La Biblioteca de Catalunya

La Biblioteca de Catalunya[11] (BC) es la biblioteca nacional de Catalunya. Fue fundada en 1907, por Enric Prat de la Riba, por lo cual el año próximo se convertirá en una institución centenaria. La BC tiene como funciones recoger, conservar y difundir el patrimonio bibliográfico de Catalunya.

La Biblioteca conserva aproximadamente 3.000.000 de documentos de diferentes tipologías y soportes, libros, revistas, manuscritos, incunables, mapas, ex libris, grabaciones sonoras, videos, material menor, CD's, DVD's y un largo etcétera.

Con un presupuesto anual el 2006 de 11'5 M de euros y una plantilla de 174 personas, ingresa cada año alrededor de 120.000 documentos nuevos por diversas procedencias, compra, donativo, intercambio y en su mayoría por depósito legal.

El año 2004, la Biblioteca de Catalunya elaboró un *Plan estratégico para 2004-2008*[12] donde se ponían de manifiesto sus prioridades, ser una biblioteca abierta, fiable y útil, y también se hacía especial énfasis en la importancia del contexto digital y en las Tecnologías de la Información, tanto desde el punto de vista instrumental para la mejora de gestión y servicios, como en tanto que soportes de patrimonio, información y conocimiento. Así pues en el plan estratégico, aprobado por su Consejo Rector -en el

[10] *Jornadas sobre Preservación del Patrimonio Digital.* (http://www.mcu.es/bibliotecas/jornadas/jppd/index.html). En las Jornadas estuvieron presentes proyectos nacionales de preservación de Holanda, Alemania, Reino Unido o Estados Unidos, entre otros paises, así como *The European Web Archive* (http://europarchive.org), la filial europea de Inernet Archive creada en 2004.

[11] Biblioteca de Catalunya. (http://www.bnc.cat)

[12] Biblioteca de Catalunya. *Pla Estratègic 2004-2008* [en línia]. Barcelona: Biblioteca de Catalunya, 2004. (http://www.bnc.cat/bc/qualitat/pestrategic2004_2008.doc)

cual está presente tanto el Consejero de Cultura del Gobierno de Catalunya como todas las administraciones locales y las instituciones y entidades representativas del sector bibliotecario- ya aparecía como uno de los objetivos estratégicos la preservación digital y la "urgencia" de comenzar a archivar la Web catalana.

En el ámbito de la preservación digital la Biblioteca de Catalunya ha puesto en marcha en los últimos dos años cooperativamente con otras bibliotecas y entidades de Catalunya diversos repositorios digitales con voluntad de proporcionar acceso permanente y garantizar la preservación patrimonial; entre los proyectos en curso destacaría *ARCA (Arxiu de Revistes Catalanes Antigues)*[13], *RACO (Revistes Catalanes amb Accés Obert)*[14]*y PADICAT (Patrimoni Digital de Catalunya)* que es el objeto de esta ponencia.

3.2 Primeros pasos

El año 2000, poco después de surgir los primeros proyectos de preservación de la Web, la BC se planteó por primera vez cómo abordar el problema. Para recoger información se contactó con los responsables del proyecto de Suecia antes mencionado, ya que su filosofía -a priori- respondía mejor a la visión que la Biblioteca tenía de la cuestión.

Estos contactos con especialistas en la preservación del patrimonio digital online, que continuaron durante el año siguiente, no pudieron concretarse en una acción planificada ya que en aquel momento la BC no disponía ni de los recursos humanos y económicos ni de la infraestructura precisa para abordar un proyecto de la magnitud que se preveía iba a tener.

En todo caso sí que sirvieron para que en la propia biblioteca se creara una sensibilidad hacia este tipo de material, se hiciera un seguimiento sobre los proyectos que progresivamente se iniciaban en Europa y se empezaran a esbozar las líneas de actuación que se deberían acometer en cuanto el contexto lo permitiera.

3.3 El diseño del proyecto

Por fin, el año 2005 pudimos iniciar la fase de diseño, la Biblioteca contó en aquel momento de un incremento del presupuesto que permitió entre otras acciones, designar un documentalista a tiempo completo.

El año 2005, fue enteramente dedicado a la recogida exhaustiva de información de primera mano, a la evaluación metódica de las diferentes opciones, a la definición de los perfiles profesionales del equipo técnico necesario para llevar a cabo el proyecto, al cálculo de costes, búsqueda de socios y financiación y test de los softwares existentes.

---

[13] ARCA (http://www.bnc.es/digital/arca/index.html) es un repositorio de acceso abierto que incluye publicacions periódicas digitalizadas (son títulos que ya no se publican, en su mayoría anteriores a 1930) representativas de la cultura y sociedad catalanas.

[14] RACO (http://www.raco.cat/) es un repositorio desde el cual se pueden consultar, en acceso abierto, los artículos a texto completo de revistas científicas, culturales y eruditas catalanas en curso de publicación.

A final de 2005 disponíamos de un informe[15] detallado de las estrategias, acciones, tareas y agentes implicados, así como de la definición del sistema de información que iba a sustentar el proyecto (software y hardware) y calendario aproximado.

El proyecto recibió por un lado el apoyo de la Generalitat de Catalunya que aprobó un presupuesto plurianual de tres años (2006-2008) dedicado a este fin, y se consiguió también un socio tecnológico –el CESCA (Centre de Supercomputació de Catalunya)[16]- con quien la Biblioteca de Catalunya ya está colaborando en otros repositorios digitales en fase de proyecto y de producción.

3.4 Qué es PADICAT

PADICAT pretende constituir un repositorio digital permanente con los recursos catalanes que se publican en la Web. Con esta finalidad se capturarán, organizaran y se proporcionará acceso a estos recursos en sus diferentes versiones, y se diseñaran los métodos y mecanismos que permitan que este patrimonio sea consultable en el futuro.

3.4.1 Estrategias de captura

PADICAT se ha diseñado como un modelo híbrido que prevé tres tipos de estrategias para identificar y capturar el patrimonio Web:

- Capturar masivamente y de forma automática los recursos Web
- Establecer convenios con los productores de los recursos
- Seleccionar Webs y recursos electrónicos vinculados a acontecimientos relevantes relacionados con Catalunya

La captura automática

En primer lugar vale la pena concretar qué entendemos por "catalán". Los organismos internacionales del sector de bibliotecas definen el patrimonio bibliográfico de un país como los documentos publicados en su territorio, en su lengua, cuya temática haga referencia al país o cuyos autores sean de dicho país. Pese a que en el contexto Web, estos conceptos no son exactamente extrapolables diremos que hemos establecido como patrimonio digital catalán:

- los recursos Web bajo el dominio .cat[17]

[15] El texto completo del informe está disponible en: Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: Biblioteca de Catalunya, desembre 2005. (http://www.bnc.cat/bc/in_padicat_2005.pdf)

[16] El CESCA (Centre de Supercomputació de Catalunya) (http://www.cesca.cat) es una corporación pública que gestiona un centro de cálculo y de comunicaciones para dar servicio a la universidad y a la investigació. La Biblioteca de Catalunya colabora ya con el CESCA en otros proyectos del ámbito digital como RACO (Revistas Catalanes amb Accés Obert) (http://www.raco.cat) y ARCA (Arxiu de Revistes Catalanes Antigues) (http://www.bnc.cat/digital/arca/index.html)

[17] El dominio .cat ha sido aprobado muy recientemente (febrero de 2006) y por lo tanto el grado de implementación es todavía bajo. Hace unos años un porcentaje elevado de los Webs de un país se situaban bajo el dominio del país (.es para España, .fr Francia, etc …), esta situación ha ido cambiando en los últimos años y ha crecido el uso de dominios no terrioriales .org,para organizaciones sin ánimo de

- los recursos Web ubicados en servidores de Catalunya
- los recursos Web en lengua catalana bajo cualquier dominio .es, .org, .edu ., .com, .inf, ...
- los recursos Web vinculados a la sociedad, lengua o cultura catalanas

Teniendo en cuenta estos criterios se capturará cualquier tipo de recurso Web. El abanico es muy amplio ya que podemos encontrar páginas Web, archivos de texto, imagen o sonido en diferentes formatos (word, pdf, excel, powerpoint, jpeg, gif, wave, mp3...), bases de datos, libros, diarios y revistas electrónicas, blogs, chats, juegos, listas de distribución, boletines de alerta y un largo etcétera.

La captura consiste en un motor de búsqueda o "spider" que recorre Internet buscando recursos Web que cumplan con los parámetros que a priori se le indican. Cada vez que identifica un recurso que responde al perfil, se capturan todos los elementos (archivos de texto, imagen, ...) que forman parte del Web y se incorporan al sistema propio.

Éste tipo de sistema automático -que es la base de los proyectos de preservación integrales o exhaustivos como el de Suecia- reciben a menudo críticas relacionadas con la "calidad" de los contenidos y su necesidad de ser preservados. En el mundo analógico la cadena de producción editorial se supone que actúa como criba de contenidos de poca calidad, pero en realidad, por poco que se conozca la producción bibliográfica, se puede observar que desgraciadamente no siempre los criterios que rigen la selección de documentos a publicar son criterios de calidad, sino que a menudo pesan más los criterios comerciales, por lo cual, los documentos impresos que se reciben en las bibliotecas por Depósito Legal tienen en su conjunto el mismo grado de interés patrimonial que los contenidos Web, ya que en un uno y otro caso, sirven para reflejar una sociedad en una época determinada a través de su capacidad de generar información, conocimiento y expresiones artísticas.

Los convenios con los productores

PADICAT es un proyecto que basará una parte importante de su éxito en conseguir que los productores y creadores de contenidos actúen como colaboradores. Cuando hablamos de productores nos referimos a un amplio abanico de agentes (administraciones, bibliotecas, museos, universidades, colegios profesionales, archivos, partidos políticos, medios de comunicación, editoriales, ...)

Si bien es cierto que en muchos casos las Webs son abiertas y accesibles, no lo es menos que en algunos, parte de los contenidos son de acceso restringido o de pago, o simplemente de uso para un colectivo determinado (es el caso de las intranets de entidades, de las hemerotecas de los diarios digitales, ...).

Esta restricción tanto puede producir que al capturar las Webs se generen Webs descontextualizadas (con enlaces huérfanos que no llevan a ninguna parte), como que las colecciones sean incompletas. Es fácil de entender si pensamos en una revista digital; las capturas automáticas requieren tiempo para que el motor y el sistema de captura recorran Internet e identifiquen y descarguen los contenidos, por lo que

---

lucro, .gov para Webs gubernamentales, .com para empresas, .edu para instituciones educativas o de enseñanza, hecho que hace más compleja la recogida del patrimonio digital.

acostumbran a poderse realizar dos, tres, o a lo sumo cuatro veces al año; si sólo nos basáramos en una estrategia de captura masiva automática, de una revista digital quincenal -por poner un ejemplo- únicamente dispondríamos de cuatro números al año de los veinticuatro editados.

Así pues la complicidad de los productores y creadores resulta imprescindible para crear una colección patrimonial de la Web completa. Básicamente, en el caso de PADICAT se solicita a los diversos agentes mediante la firma de un convenio que autoricen a la Biblioteca a capturar, tratar y difundir las diferentes versiones de sus Webs.

Seleccionar Webs y recursos electrónicos vinculados a acontecimientos relevantes relacionados con Catalunya

La tercera estrategia contribuye a dotar de valor a la colección patrimonial Web. Se trata de identificar y capturar de forma exhaustiva todos los contenidos digitales relacionados con un hecho destacado de carácter político, cultural, social o deportivo. Son acontecimientos que normalmente se desarrollan en un periodo limitado de tiempo pero generan mucha información Web de opinión, presentación o propaganda.

Complementariamente la Biblioteca de Catalunya, así como el conjunto del ámbito bibliotecario, continua reclamando la necesidad de redactar una nueva ley sobre el depósito legal que incluya los recursos Web, y regule tanto su captura como su difusión y tratamiento, aspectos éstos dos últimos para los que el vacío legal actual es abrumador.

3.4.2 Organización y recuperación de los recursos Web

Una vez capturados los recursos Web, el sistema los organiza, identifica, indexa a texto completo y los guarda en la base de datos (en formato de compresión .arc), a punto para ser recuperables.

La búsqueda y recuperación de la información se puede realizar por la dirección URL, por el título de la Web, o mediante texto completo; esta última, sin duda una buena opción para recuperar la información sistema “google”, genera, sin embargo, para búsquedas muy genéricas, demasiado ruido y conjuntos enormes de resultados[18]; es necesario que la ordenación de los resultados sea por su pertinencia.
En el contexto del proyecto PADICAT, estamos adaptando el sistema para incrementar la eficacia de la búsqueda; optaremos por catalogar mediante metadatos Dublin Core[19] los recursos seleccionados que se consideren de interés elevado desde el punto de vista patrimonial, de esta manera podremos ordenar los resultados de forma que los usuarios obtengan en primer lugar los más pertinentes y a continuación todo el resto de Webs encontrados fruto de la búsqueda a texto completo.

[18] Tengamos en cuenta la magnitud de datos que se gestionan; por ejemplo, el proyecto sueco (exhaustivo) ha capturado hasta 2005, 305,85 millones de archivos (*Kulturarw3. Statistik.* http://www.kb.se/kw3/Statistik.htm), pero es que incluso el proyecto australiano (selectivo) maneja 31,5 millones de archivos (*PANDORA archive size and monthly growth* (http://pandora.nla.gov.au/statistics.html)

[19] La catalogación por metadatos Dublin Core (http://dublincore.org) consiste en la creación de un registro descriptivo simple del recurso con quince elementos informativos (autor, título, lengua, ...); estos metadatos facilitan la identificación de los recursos por parte de los motores o buscadores.

3.4.3 Difusión pública

Una de las cuestiones a resolver en este tipo de proyectos es la difusión Web. La legislación sobre difusión web es aún ambigua, no queda pues claro qué se puede o no se puede hacer. Los aspectos a considerar son diversos: el acceso en Internet o en una red local, la posibilidad de obtener impresión o la descarga de ficheros.

En el contexto de Catalunya, nuestra visión y voluntad es hacer público en Internet el contenido del repositorio, de aquí parte de la importancia de conseguir la colaboración de las editoriales y productores de contenidos digitales en Web.

La voluntad de la Biblioteca de dar acceso abierto al repositorio requiere de un sistema eficiente que en el caso de existencia explícita de una negativa por parte de un productor de publicar sus contenidos, éstos puedan ser rápidamente bloqueados de la difusión en abierto, y restringidos únicamente a la consulta dentro de la propia red de la Biblioteca de Catalunya.

Los convenios que hemos empezado a distribuir entre productores de Catalunya incluyen la autorización explicita de difundir los contenidos en Internet. Cabe considerar que por la forma de presentación de los contenidos PADICAT no constituye una competencia para las editoriales o agentes. Se trata de una presentación cronológica de las diferentes versiones de una misma Web o recurso Web, que incluso cuando se trate de contenidos de pago, pueden establecerse los "embargos"[20] que se acuerden con el editor o productor de la publicación.

3.4.4 Acciones y estrategias de preservación

La preservación del patrimonio digital Web debe afrontar dos problemáticas complementarias: la perdurabilidad del soporte físico y la obsolescencia del software y del hardware.

Los soportes en los que se graban y almacena la información digital tienen una previsión de vida relativamente corta que hace necesario planificar actuaciones a medio y largo plazo para garantizar el acceso permanente. Hoy en día aún no se ha desarrollado ningún soporte que asegure una perdurabilidad similar a la del papel (al menos 500 años). Los soportes magnéticos tienen solo una "durabilidad" de entre 3 y 30 años y los ópticos de 5 a 100 años.

La solución que temporalmente están adoptando las instituciones es el "refreshing" o actualización, que consiste en la transferencia de datos a un nuevo soporte físico sin aplicar ningún tipo de conversión o modificación en su estructura lógica.

A la problemática de los soportes se añade la obsolescencia del software y del hardware: aparición de nuevas versiones del software, a veces incompatibles o parcialmente compatibles con las versiones anteriores, nuevos entornos y plataformas.

[20] El uso de "embargos" es bastante frecuente en la publicación de revistas electrónicas en abierto. Consiste en mantener bloqueado el acceso a los números más recientes de la publicación y dar en abierto el resto de la colección.

Entre las posibles soluciones ante esta casuística encontramos la emulación, la migración o la encapsulación.

La migración presenta a menudo problemas de pérdida de las funcionalidades del entorno original en el proceso de transferencia o conversión de la información hacia un sistema informático diferente.

La emulación o capacidad de que los sistemas informáticos futuros puedan recrear las funcionalidades y comportamiento originales de los documentos requiere de un compromiso de las empresas de software del futuro difícil de garantizar; cabe destacar que aunque difícilmente se pueda optar por esta metodología, la emulación seria probablemente la mejor solución para conservar la integridad de los documentos digitales, es decir para poder a largo plazo consultarlos tal como fueron creados en origen (datos, funcionalidades y aspecto).

La encapsulación permite, en esencia, conocer toda la información necesaria para dar acceso a un objeto digital (por ejemplo: sus componentes, las relaciones entre ellos, como interpretar los bits digitales)

Finalmente, una posible solución, que ha sido ya prácticamente descartada consistiría en conservar el hardware y software originales, lo que convertiría las bibliotecas en museos informáticos sometidos a un mantenimiento constante.

Es aún prematuro determinar las acciones de preservación que se deberán aplicar en el futuro para garantizar el acceso permanente a los contenidos del repositorio. Existe bibliografía abundante[21] y son muchas las instituciones y organizaciones involucradas en buscar soluciones. Se trata pues de estar involucrados en estos forum y trabajar cooperativamente con otros proyectos y países para determinar las acciones a ejecutar.
En los proyectos de preservación de la Web un aspecto favorable es que más del 96 % de la información se encuentra en formatos estándares (texto/html, imágenes en jpeg, gif o pdf.), por lo que las soluciones que se adopten lo serán para prácticamente todo el repositorio. Para el 4% restante (formatos y aplicaciones efímeros, softwares propietarios, ...) se deberá buscar soluciones a la medida de cada caso.

3.4.5 Infraestructura

El repositorio PADICAT está situado físicamente en la sede del CESCA, nuestro socio tecnológico. El hardware que soporta el sistema consiste en:

- dos servidores de dos nodos ProLiant DL360 G4p,
- el clúster de e-información,
- el robot Scalar i2000
- 10 TB[22] de espacio para almacenamiento de datos, que se prevé pueda ampliarse anualmente con 10 TB adicionales.

[21] Muy recomendable la bibliografia y proyectos sobre preservación digital en PADI (http://www.nla.gov.au/padi/)

[22] 1 TB (Terabyte)=1.024 GB (Gigabyte)= 1.048.576 MB (Megabyte)

El software, de código libre y gratuito, es el conocido como Heritrix y que agrupa una serie de softwares que gestionan las diferentes partes del sistema de gestión:

- Heritrix, el robot de captura (*crawler)*
- BAT (*Bnf Arc Tools*), el gestor de archivos
- NutchWax (*Nutch Web Archive eXtensions),* el buscador en el propio sistema
- Wera. *(Web Archive Access)* la interfaz de consulta

3.4.6 Equipo técnico

Lógicamente se trata de un equipo multidisciplinar que cuenta además con el soporte y asesoramiento de los especialistas de la propia Biblioteca de Catalunya en materias como definición de los metadatos, aspectos legales, informática, usabilidad, ...

A día de hoy están trabajando en el proyecto el Coordinador, con un perfil de documentalista, y el Analista informático. El resto del equipo se irá incorporando paulatinamente en función de la evolución del propio proyecto a partir del último trimestre de 2006 y a lo largo del 2007.

Cuando esté completamente constituido estará formado por 7 técnicos de diferentes perfiles:

- Coordinador del proyecto
- Analista informático
- 2 Bibliotecarios de soporte
- 2 Bibliotecarios especialistas en metadatos
- 1 Administrativo (para la gestión de trámites y convenios)

3.4.7 Estado del proyecto y previsiones futuras

De acuerdo con el calendario establecido en la fase de diseño a julio de 2006 se han completado las siguientes acciones:

- Instalación, adaptación y parametrización del software para adecuarlo al proyecto. Posteriormente en función de los resultados y rendimiento se procederá a introducir los ajustes o modificaciones que se consideren oportunos.
- Establecimiento de los procedimientos y estándares para la identificación de agentes, validación y catalogación de recursos Web.
- Inicio de una primera captura de recursos sobre la base de un conjunto controlado y limitado de 100 URL's.
- Adaptación de la interfaz de consulta con la previsión de poder hacerla pública de forma experimental a partir del mes de septiembre de 2006.
- Identificación de 140 instituciones catalanas representativas de la administración pública, sociedad civil y medios de comunicación, con las que ya se ha establecido un primer contacto para que sean socios colaboradores del proyecto (en poco tiempo –cuestión de días- hemos recibido respuesta de un 25% de ellas con una percepción positiva de PADICAT).

Los objetivos de PADICAT para el año 2009 son:

- 100.000 recursos Webs formando parte del repositorio
- 50 millones de archivos
- 30 TB de volumen
- Acuerdos con 300 instituciones de todo tipo
- Acceso en línea a buena parte de la colección

Conclusiones

El inicio de acciones de preservación y acceso permanente al patrimonio digital es una medida necesaria que todos los países deben impulsar. La Web crece exponencialmente y a diario surgen y desaparecen contenidos que pueden ser significativos para reproducir la historia de una sociedad.

La preservación de la Web se debe impulsar desde las administraciones, de la misma manera que durante años han habilitado los mecanismos para la conservación del patrimonio analógico y han dado apoyo -en mayor o menor medida- a las instituciones de la memoria.

Por sus características, complejidad y magnitud la preservación del patrimonio digital debe hacerse de forma cooperativa, generando alianzas y convergencia de intereses, y distribuyendo responsabilidades de manera que resulte beneficiosa para todos los agentes implicados.

En el contexto actual, con numerosas iniciativas en curso en otros países, el conocimiento y experiencia que de ellos se pueda obtener constituye una gran ventaja que permite reducir errores y potenciar las buenas prácticas.

El proyecto PADICAT de Cataluña surge de un estudio y análisis detallado del estado de la cuestión a nivel mundial lo que ha permitido alinearse en la tendencia generalizada en cuanto al modelo a seguir. A su vez, PADICAT es un proyecto de preservación de Web territorial, que puede en el futuro actuar de referente para escenarios similares de países sin dominio propio.

Queda mucho camino por recorrer pero por fin, con el proyecto PADICAT, en Cataluña hemos dado un primer paso que esperamos y confiamos que sea un camino que tenga como futuro la consolidación y la mejora del sistema, para garantizar a las generaciones futuras el conocimiento y conservación de su herencia cultural