bid textos universitaris de biblioteconomia i documentació
número 15
desembre de 2005
ISSN 1575-5886
DL B-19.675-1998
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales

[Versió catalana]

CIRO LLUECA FONOLLOSA
Projecte PADICAT (Patrimoni Digital de Catalunya)
Biblioteca de Catalunya
cllueca@bnc.es

Resumen [Abstract] [Resum]

Las tecnologías de la información y la comunicación han facilitado que el patrimonio cultural, científico y la información en general se presenten en formato digital, así como en los formatos analógicos tradicionales. La reacción no se ha hecho esperar, y desde la década de los noventa han surgido diversos proyectos destinados a garantizar el acceso permanente a la producción digital —su recopilación y almacenaje, el tratamiento, la preservación y la difusión—. Se presenta la panorámica mundial de los modelos existentes de *depósitos digitales nacionales*, nombre que reciben estos proyectos impulsados habitualmente por las bibliotecas nacionales, con un objetivo común: hacer que las páginas web sean siempre accesibles.

1 Introducción

Las tecnologías de la información y la comunicación han facilitado que el patrimonio cultural, científico y la información en general se presenten en formato digital, así como en los formatos analógicos tradicionales. Actualmente, tal como recomiendan las *Directrices para la preservación del patrimonio digital*,[1] los recursos que son fruto del conocimiento o la expresión de los seres humanos, ya sean de carácter cultural, educativo, científico o administrativo, o comprendan información técnica, jurídica, médica y de otro tipo, se generan directamente en formato digital o se convierten a este formato a partir de material analógico ya existente. Los productos que se generan digitalmente no existen en otro formato que no sea el electrónico original.

Esta realidad, sumada a la voluntad de las personas, las instituciones y los gobiernos de velar por la preservación de cualquier forma de patrimonio, ha posibilitado que las administraciones de diversos países hayan iniciado políticas destinadas a garantizar el acceso permanente a la producción digital —su recopilación y almacenaje, el tratamiento, la preservación y la difusión—, por parte de los agentes públicos y privados.

Las dificultades son notables. Para empezar, los métodos tradicionales de preservación de la producción bibliográfica (como el depósito legal) son de difícil aplicación en el entorno digital porque, a parte de la posible obsolescencia del texto legal (el caso español), los recursos digitales pueden instalarse en servidores de cualquier lugar del mundo (como pasa también con los impresores que no son editores). Este hecho dificulta la tradicional correspondencia geográfica entre la ubicación del productor y la lengua o la temática publicada. En segundo lugar, la producción digital tiene un crecimiento exponencial, siendo además muy variable la durabilidad de los materiales publicados en Internet[2] y, en consecuencia, limitada la posibilidad de acceso permanente al patrimonio. Finalmente, señalar la cuestión de la propiedad intelectual del producto digital, sin un derecho basado en el principio de copia para la preservación que asegure la conservación

y perdurabilidad del patrimonio digital, con las limitaciones comerciales que sean necesarias.

Como se ha apuntado, y pese a las dificultades, diversos países han entendido la necesidad de pasar a la acción, y de establecer políticas y emprender acciones de preservación para asegurar la pervivencia de la producción digital, como ya se había hecho históricamente con los documentos impresos y en soportes tradicionales, mediante las leyes nacionales del depósito legal. En la mayor parte de los casos que se presentarán ha sido la biblioteca nacional quien ha liderado el proceso de preservación y el acceso al patrimonio digital; para hacerlo ha implicado al resto de agentes.

En junio de 2005 la Biblioteca de Catalunya puso en marcha el proyecto PADICAT (Patrimoni Digital de Catalunya. El texto que ahora se presenta es uno de los frutos de la primera fase del proyecto. Es objetivo de este artículo presentar la panorámica mundial en materia de accesibilidad permanente a la producción web,[3] y los modelos existentes para capturar las producciones digitales en línea. No se examinan otras cuestiones de estos modelos, como los aspectos legales de cada territorio o las diferencias de los programas informáticos utilizados por las bibliotecas.

2 Depósitos digitales nacionales: *archivando la web*

En los orígenes de la preservación digital podemos mencionar las acciones de las bibliotecas virtuales con proyectos de bibliotecas de investigación, universitarias, nacionales, y también públicas, dedicados a presentar directorios temáticos de recursos electrónicos. Complementariamente y para dar respuesta a necesidades temáticas concretas (normalmente de ámbito geográfico que seguían el modelo enciclopédico digital de los CD-ROMs) se optó por crear depósitos multiformato (imágenes, sonido, texto, gráficos, etc.). El siguiente paso ha sido conservar los recursos propios para garantizar el acceso con todas las variables formales que se han producido en el tiempo: son las iniciativas denominadas *depósitos institucionales, archivos de e-prints*, etc. Cuando el proceso se dedica a un territorio, hablamos de los *depósitos*[4] *digitales nacionales, archivos web*, o *bibliotecas nacionales digitales*.

Un depósito digital nacional tiene la misión de garantizar el acceso a largo plazo a los recursos digitales que se generan en un territorio, o sobre un territorio determinado. De hecho, la misión de la Biblioteca de Catalunya, como biblioteca nacional, no es otra que 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. Y entendemos que este patrimonio bibliográfico incluye también la producción bibliográfica digital catalana que conformará el PADICAT, el Patrimonio Digital de Catalunya.

2.1 Los modelos existentes

Las experiencias existentes de depósitos nacionales digitales se agrupan en dos modelos. Por un lado, el modelo integral o exhaustivo (modelo mayoritario, y característico de Suecia, Noruega, Finlandia, Islandia y Austria, entre otros) que apuesta por la integración automática del total de la web objeto de preservación a partir de determinados criterios infraestructurales (lingüísticos, según el dominio de las webs, según la ubicación del servidor, etc.). Por otro lado, el modelo selectivo (seguido por Australia, Canadá, Japón y el Reino Unido entre otros países) que dirige las acciones de recopilación de acuerdo con una política selectiva temática sobre un espacio geográfico determinado, sobre un tema de interés nacional, etc., mediante acuerdos con los editores o productores de recursos web.

Estos dos modelos han dado paso en algunos países, y cada vez con más fuerza, a modelos híbridos que complementan la captura periódica del total de la web nacional con

acuerdos con los productores, que se basan en intereses temáticos o que tienen relación con acontecimientos de actualidad (elecciones, catástrofes, etc.).

Finalmente, otros análisis teóricos[5] de la situación apuntan a una clasificación más compleja, según si la web a capturar es estática o dinámica, por ejemplo, pero entendemos aquí que la captura restrictiva de la web según su complejidad de preservación y garantía de acceso,si es estática, es más fácil capturarla y preservarla, es sólo un primer paso hacia el objetivo final de todos los depósitos digitales nacionales.

2.1.1 El modelo integral

Las principales ventajas o puntos fuertes del modelo integral son:

- *Riqueza de la colección, en cantidad y en calidad*. De acuerdo con este modelo, no se condiciona selectivamente qué es interesante y qué no lo es, sobretodo si se considera que difícilmente somos capaces de prever cuáles serán los usos y las líneas de investigación futuras. Al reflejar, además, el crecimiento de la web y los cambios en el diseño de la publicación web a todos los niveles, se aporta un componente sociológico añadido, ya que las páginas personales, los *weblogs*, los chats, e *in extremis* los videojuegos en línea, forman parte también de la producción digital nacional en los depósitos integrales. Ligado a este punto fuerte está el hecho de que la captura exhaustiva permite respetar en buena medida la interrelación de las sedes web.

  A título de ejemplo, el proyecto sueco, Kulturarw3, contiene las sedes web con dominio .se, .nu,[6] las sedes web con dominios internacionales (.com, .org, .net) ubicadas en servidores en territorio sueco[7] y la *Suecana extreana* —las webs que tratan sobre Suecia, viajes por Suecia, o traducciones de obras literarias suecas.

- *Compilación automática*. El hardware y software utilizados aseguran una alta capacidad de captura y de almacenaje en respuesta a unos parámetros determinados, que pueden ser tan amplios como la dirección del proyecto establezca. Este hecho minimiza los recursos más costosos, los de personal, a la vez que consigue resultados visibles (presentables y vendibles a la sensibilidad política y ciudadana) en un período de tiempo razonablemente corto, pues en poco tiempo se crea una colección significativa.

  Finalizada la primera captura del proyecto finlandés (agosto–septiembre de 2001), los responsables de la Helsingin Yliopiston Kirjasto-Suomen Kansalliskirjasto aseguran que se han capturado 7,5 millones de URL, y calculan que esta fotografía de la web finlandesa representa entre el 30 y el 50% del total capturable.

- *Bajo coste*. Con relación a los aspectos ya descritos anteriormente, se observa que los proyectos que apuestan por un modelo integral están coordinados por equipos de pocas personas. Pese a que estos equipos están integrados en las estructuras de las bibliotecas nacionales, lo cierto es que habitualmente no se dedican prácticamente recursos de personal a la catalogación y gestión de los procesos selectivos.

  Mientras que el proyecto australiano Pandora (selectivo) dedica un total de trece personas a tiempo completo y el sistema utilizado en Quebec (selectivo e integrado al catálogo IRIS) tiene un equipo de diez personas, los proyectos basados en la captura integral tienen equipos sensiblemente menores —por ejemplo, los proyectos de Austria y Suecia tienen equipos de una a tres personas—.

Los principales inconvenientes o puntos débiles del modelo integral son:

- *Imposibilidad de acceder a la Internet invisible*. El sistema automático de captura sólo tiene acceso a los recursos que están publicados en abierto, y por tanto se producen lagunas en la captura de webs de pago, webs protegidas con contraseñas, páginas huérfanas, y la mayor parte de les páginas dinámicas. Asimismo tampoco es posible acceder a las infranet o bases de datos bibliográficas (catálogos de bibliotecas) o alfanuméricas (diccionarios). la Internet invisible multiplica la Internet visible entre dos y cincuenta veces. Además, este espacio supone un archipiélago de calidad por el tipo de recursos que contiene (artículos, estudios científicos, publicaciones digitales, etc.).

  El hecho es que, como han explicado repetidamente Isidro Aguillo y los investigadores del InternetLab,[8] la Internet invisible multiplica la Internet visible entre dos y cincuenta veces. Además, este espacio supone un archipiélago de calidad por el tipo de recursos que contiene (artículos, estudios científicos, publicaciones digitales, etc.).

- *Compilación irregular de la colección*. Esta característica es consecuencia de las lagunas en el control de los ítems coleccionados (por ejemplo, en las publicaciones periódicas), la no reclamación de los documentos no accesibles, la pérdida de documentos importantes y de los cambios frecuentes que se producen en determinadas sedes web.

  En esta línea, el proyecto noruego Paradigma comenzó sus capturas integrales en 2001 con el dominio .no. En la tercera captura (agosto de 2003) se ha ampliado el alcance a los dominios internacionales (.com, .net) y a 65 diarios digitales noruegos que quedaban excluidos de las capturas periódicas. Es un ejemplo de un modelo integral con acciones selectivas.

- *Acceso limitado a los resultados*. La falta de un proceso de catalogación (metadatos, inclusión en el catálogo de la biblioteca, etc.), dificulta la recuperación de los documentos capturados. Por otra parte, el respeto a los derechos de autor para publicar sin autorización o sin acuerdo previo los recursos capturados comporta que se restrinja el acceso a los depósitos nacionales y que normalmente sólo se puedan consultar desde las propias instalaciones de las bibliotecas nacionales Esta medida no acaba de conjugar con la propia naturaleza de los proyectos: garantizar el acceso a la producción digital.

  De los proyectos con intención exhaustiva, únicamente Internet Archive ofrece acceso abierto y en línea a sus fondos, y actualmente sólo permite la búsqueda por URL. El resto de proyectos —incluidos los escandinavos, que son los más veteranos— limitan la consulta de sus colecciones a las dependencias de las bibliotecas nacionales que los lideran. Sólo una parte muy pequeña de la colección recibe un trato que facilita la recuperación más allá del URL o la fecha de captura.

2.1.2 El modelo selectivo

Las principales ventajas o puntos fuertes del modelo selectivo son:

- *Creación de una colección equilibrada*. Cada ítem que formará parte del archivo es evaluado teniendo en cuenta su pauta de publicación. El modelo responde, pues, a pautas más cercanas al modelo bibliotecario clásico, en el sentido que se conoce lo que forma parte de la colección y que ésta se amplia teniendo en cuenta la realidad del territorio y de todos los usuarios.

  Por ejemplo, el modelo de crecimiento de Pandora es visible también en el Reino Unido, con el UKWA (United Kingdom Web Archive). El recurso presenta una clasificación de los contenidos muy similar al popular directorio *Yahoo*. A partir de

un máximo de nueve categorías iniciales (Arte y humanidades, Negocios y economía, etc.) se produce una taxonomía en cascada (Arte y humanidades: Arquitectura; Danza; Bellas artes; Geografía; Historia; Lenguas; Literatura; Música...) y la presencia de los recursos digitales británicos es temáticamente equilibrada en todas ellas.

- *Máxima facilidad de acceso al fondo.* Cada ítem puede ser completamente catalogado y pasar a formar parte de la bibliografía nacional, de manera que los datos bibliográficos pueden compartirse. Los recursos se integran en el catálogo de la biblioteca, y los acuerdos permiten publicar los recursos en línea, abiertamente. La catalogación de los documentos hace que las posibilidades de recuperación sean ilimitadas. Prueba de ello es que en junio de 2004, el proyecto WARP (Japón) facilitaba el acceso completo a 600 sedes web (administración, universidades, congresos y seminarios) y a 110 diarios electrónicos. Lituania tiene su archivo de recursos electrónicos integrado en el catálogo colectivo LIBIS. Finalmente, Pandora (Australia) tiene integrado su fondo en el catálogo de la biblioteca, y permite a los buscadores (*Google, Msn*, etc.) acceder a los recursos desde determinados niveles o parámetros de búsqueda.

- *Estratégico*. Al funcionar con alianzas y acuerdos con las entidades editoras (comerciales o no) la implicación de los agentes productores se produce más naturalmente, con voluntad compartida. Además, los acuerdos hacen posible que los ítems sean accesibles en línea en toda su extensión, al tiempo que la información formal y propiedades de cada recurso son conocidos por los gestores de captura. El modelo permite desarrollar métodos y herramientas de compilación y acceso y estrategias de preservación a más largo plazo. La web invisible y la infranet se incluyen en el depósito.

  Posiblemente sea Pandora el ejemplo más evidente, pero no el único, de la fuerza de la cooperación. La National Library of Australia cuenta con diversas instituciones socias del proyecto: el Australian War Memorial, el Australian Institute of Aboriginal and Torres Strait Islander Studies y las bibliotecas de los diferentes estado del país entre otras. La alianza proporciona rigor en la selección, soporte a los presupuestos y presencia mediática y en la comunidad científica australiana.

Los principales inconvenientes o puntos débiles del modelo selectivo son:

- *Parcialidad al describir el mundo*. En la selección de los recursos se realiza un juicio subjetivo sobre su valor y se anticipa aquello que los investigadores preferirán en el futuro. En todo caso, la extensión de un archivo selectivo es muy limitada en comparación con el volumen del material de un territorio determinado, y pese a los esfuerzos, los criterios de selección son de difícil definición.

  El proyecto británico UKWA está liderado por la British Library y cuenta, entre otros socios importantes, con la National Library of Wales y la National Library of Scotland. El proyecto está en desarrollo, y los primeros resultados se han hecho públicos recientemente (mayo de 2005), pero el hecho es que algunos ítems seleccionado por los socios de proyecto dan una visión parcial de la web británica. Por ejemplo, bajo el epígrafe “teatro” el único ítem disponible es “theatre in Wales”.

- *Coste elevado*. La selección, la gestión y el seguimiento de los acuerdos y las capturas, y especialmente el análisis documental de los recursos son tareas muy intensivas que encarecen el coste por ítem en recursos humanos. El hecho de que las instituciones que gestionan los depósitos sean las bibliotecas nacionales garantiza una alta calidad en la descripción e indización de los recursos,

habitualmente por lenguaje de metadatos.

En el congreso celebrado en Canberra en noviembre de 2004,[9] la responsable del proyecto australiano Pandora desvelaba que el coste por la gestión de un ítem digital puede llegar a ser cinco veces superior al de una monografía.

- *Descontextualización de la colección*. La selección de los recursos no se realiza necesariamente en su contexto, y por tanto no incluye los recursos enlazados que contextualizan la información. En el lenguaje del hipertexto la selección de una web sin tener en cuenta con qué otras está enlazada puede dar una lectura huérfana del recurso.

  El problema principal del emblemático Pandora es la preservación de los links rotos. La política que se sigue es dar la posibilidad al usuario de acceder a la versión actual de la web a la cual apuntaba el link del recurso preservado, pero los problemas de contexto no quedan resueltos.

2.1.3 El modelo híbrido

Los dos modelos anteriores han dejado paso a modelos híbridos que complementan la captura sistemática de la web nacional (modelo típicamente integral) con acuerdos con instituciones productoras según los intereses temáticos (modelo selectivo). Adicionalmente, el proyecto se puede dirigir a efectuar capturas selectivas de determinados acontecimientos de interés general, como por ejemplo juegos olímpicos, elecciones, catástrofes naturales, etc.

El caso danés, por ejemplo, está enfocado en la acción triple que supone la captura exhaustiva de la web danesa, los acuerdos con entidades editoras del país y la captura exhaustiva pero focalizada de acontecimientos de interés.

Del estudio detallado de los depósitos existentes se desprende que esta es la tendencia a seguir por la mayoría de los proyectos integrales (Austria, Países Bajos, Suecia, Finlandia, etc.).

Lógicamente, los proyectos híbridos incorporan algunas de las ventajas descritas anteriormente (colección rica y equilibrada, máximo acceso, impulso de los acuerdos estratégicos, compilación automatizada y seguimiento de las lagunas), pero también elementos negativos (coste elevado, equipos más numerosos, carga de gestión).

2.2 Depósitos digitales nacionales

Se han hallado referencias de veinte casos de depósitos nacionales:[10] Alemania, Australia, Austria, Canadá, Dinamarca, Estados Unidos de América, Estonia, Finlandia, Francia, Grecia, Islandia, Japón, Lituania, Noruega, Nueva Zelanda, Países bajos, Québec, Reino Unido, República Checa y Suecia. Atendiendo al modelo que sigan, se pueden clasificar de la manera siguiente:

- Modelo integral (50%): Alemania, Austria, Estonia, Finlandia, Grecia, Islandia, Lituania, Noruega, República Checa y Suecia.
- Modelo selectivo (35%): Australia, Canadá, Estados Unidos de América, Japón, Países bajos, Québec y Reino Unido.
- Modelo híbrido (15%): los depósitos de Dinamarca, Francia y Nueva Zelanda se puede considerar proyectos que se enmarcan plenamente en el modelo híbrido.

Sin embargo, como ya se ha mencionado, la mayor parte de los depósitos que siguen un modelo integral han adoptado medidas para incluir determinados recursos (como publicaciones periódicas) que los acercan a parámetros híbridos. Esta es la tendencia

generalizada.

Analizaremos a continuación los casos existentes, centrándonos en tres ejemplos que representan los modelos anunciados: integral (Kulturarw3 de Suecia), selectivo (Pandora de Australia) e híbrido (Netarkivet de Dinamarca). Del resto de proyectos se hace una descripción somera de características en el anexo. No se aportan datos de Islandia ni de Estonia por falta de bibliografía.

2.2.1 Kulturarw3 (http://www.kb.se/kw3/ENG)

El proyecto está liderado por la Kung. Royalbiblioteket (Suecia). Se inicia en el año 1996 y sigue el modelo integral. Es exhaustivo por lo que afecta a la captura de la web sueca —350.000 webs (febrero de 2005)— y la catalogación de los materiales no es una prioridad. El acceso al fondo está limitado a las dependencias de la Kung. Royalbiblioteket.

El caso sueco es un paradigma de anticipación. A partir de los orígenes del depósito legal, de 1661, la revisión de la legislación que se lleva a cabo en 1993 incluye la información electrónica publicada en soportes tangibles (archivos informáticos, CD-ROM). La biblioteca sueca crea en 1996 el Kulturarw3, el archivo web sueco. Seis años más tarde, en 2002, se decreta en Suecia que la biblioteca nacional realiza *de iure* los trabajos de preservación y accesibilidad permanente del patrimonio digital sueco.

La colección cubre revistas digitales y publicaciones periódicas no diarias, a excepción, desde hace unos meses, de una selección de más de 100 títulos de diarios suecos, documentos estáticos (archivos electrónicos) y documentos dinámicos con enlaces. Ulteriormente se recopila el contenido de listas de discusión, y archivos FTP abiertos. Por los datos hechos públicos en febrero de 2005, sabemos que en aquella fecha las dimensiones de Kulturarw3 eran de 306 millones de archivos y unos 10.000 Gb: 350.000 sedes web.

Las herramientas de captura y organización son el programa Combine y, más recientemente para los diarios digitales, Heritrix.

Los puntos fuertes del proyecto Kulturarw3 son los derivados del exhaustivo de la compilación automática y el valor en plasmar la sociedad digital sueca.[11] Se pueden hallar desde revistas científicas a *weblogs* de ONG. Los puntos débiles están relacionados con importantes lagunas en el control de aquello que se captura, en la nula profundidad en la Infranet (contenidos de pago o con contraseña, páginas huérfanas, etc.) y la falta de catalogación del archivo. Como se ha apuntado, el hecho de que el archivo sea únicamente consultable en les dependencias de la biblioteca sueca es una característica negativa del sistema.

2.2.2 Pandora (http://Pandora.nla.gov.au/index.html)

El proyecto está liderado por la National Library of Australia (Australia). Se inicia en el año 1996 y sigue el modelo selectivo. Su alcance se centra en la selección de publicaciones en línea y webs sobre Australia, de autor australiano o sobre tema australiano. La catalogación es exhaustiva y les posibilidades de búsqueda, muy avanzadas. Dispone de un software propio, *Pandas*, que se ha implementado en otros proyectos.

El archivo web de Australia, Pandora, fue creado en 1996 por la National Library of Australia para garantizar el acceso permanente a una selección de publicaciones en línea y sedes web de y sobre Australia.

A falta de una ley que regule el depósito legal digital (la vigente es de 1968), la política

de la biblioteca y los socios de proyecto, que forman el comité científico de la política selectiva, es llegar a acuerdos con las entidades editoras de los documentos susceptibles de ser capturados. Existe una guía publicada con los criterios de selección de las sedes capturadas. Los datos estadísticos de septiembre de 2005 muestran que el archivo contiene 27 millones de ficheros y tiene un crecimiento mensual de 30 Gb. Es consultable en línea.

Los inconvenientes del sistema australiano están relacionados con su propia naturaleza:[12] el criterio de la selección es forzosamente subjetivo, pese a la transparencia de la política de selección. El contexto (los enlaces a los cuales apunta el recurso), quedan desligados del documento, porque pueden no estar incluidos en la selección. Finalmente, el coste de tratamiento (selección, captura periódica, catalogación, etc.) de cada ítem es muy elevado.

Por contra, los beneficios se concentran en la calidad del tratamiento y la presentación del patrimonio. La accesibilidad en línea, en abierto, es posible por los acuerdos subscritos con los productores (que comporta el acceso a los recursos de la infranet). Los datos de la catalogación son compartibles con el resto de equipamientos australianos (o internacionales). Se procura un crecimiento temático equilibrado de la colección.

Vistos los modelos integral y selectivo, la tercera vía a considerar es la mixta. Como se ha mencionado, buena parte de los depósitos digitales nacionales planteados inicialmente como integrales han ido adoptando medidas para incluir recursos muy significativos, como publicaciones periódicas, en sus fondos.

Son tres los proyectos pioneros en apostar por una política clara de conjugación de las capturas exhaustivas, los acuerdos con instituciones y organizaciones, y el detalle para actividades concretas: Francia, Dinamarca, y Nueva Zelanda. A continuación se examina con detalle el proyecto danés.

2.2.3 Netarkivet (http://netarchive.dk/index-en.php)

El proyecto está liderado por Det Kongelige Bibliotek (Dinamarca). Se inicia en el año 1998 y sigue el modelo híbrido, que se basa en la captura exhaustiva, los acuerdos con las instituciones y las actividades especiales relacionadas con la realidad danesa. Los trabajos de captura integral se han iniciado en julio de 2005. Desde 1997 la ley de depósito legal incluye “todas” las publicaciones de Dinamarca, y desde 2005 la biblioteca danesa tiene potestad para capturar todo tipo de páginas web danesas. La Kongelige Bibliotek facilita la entrega del depósito legal por medio de un formulario web.

A partir de un modelo inicial (dominio .dk), en 2004 se adopta el sistema híbrido, dirigido como se ha mencionado al triple objetivo (integral+selectivo+especiales). En él participan la Kongelige Bibliotek (la biblioteca nacional de Dinamarca) y la State and University Library (ubicada en Uhus). Puntualmente (en la selección de sedes web de literatura danesa, por ejemplo) se incorporan entidades vinculadas temáticamente.

El 24 de junio[13] de 2005 se anuncia al gran público la puesta en marcha del proyecto danés en su vertiente integral, ligándolo a la nueva modificación de la Ley de depósito legal.

Después de la primera fase de captura integral de julio de 2005 con el software *Heritrix*, el proyecto Netarkivet contiene 600.000 dominios .dk, cifra que representa, según la propia institución, el 60% de los dominios daneses. Paralelamente, hasta el año 2004 y por el modelo selectivo, el depósito ronda los 500 Terabytes de volumen, con un exponente de crecimiento anual de 30 Tb.

No es accesible en línea.

El proyecto danés tiene puntos fuertes evidentes que son infraestructurales:

- Dinamarca es un país pequeño (5,4 millones de habitantes), con un dominio propio (dk) y lengua propia (danés), la selección (y las gestiones consecuentes) es relativamente sencilla porque es limitada.
- Por la misma razón, son también aparentemente sencillas las capturas exhaustivas. Para el desarrollo del software de captura selectiva y gestión, el Netarkivet ha colaborado con el Nordic Web Archive (NWA), junta al resto de bibliotecas escandinavas.
- La legislación que afecta al depósito legal ha ido modificándose (la última revisión es principios de julio de 2005),[14] para ampliarse a las publicaciones digitales en línea. Un formulario en línea facilita la labor de productores y editores web.
- Existen dos socios importantes involucrados en la coordinación del proyecto: la biblioteca nacional y el centro de investigación de una importante universidad. Se elaboran acuerdos estratégicos en función de determinadas temáticas.

2.3 Organizaciones y proyectos suprainstitucionales

Los proyectos descritos, que llevan a cabo las bibliotecas nacionales, se encuentran a menudo bajo "paraguas" más amplios de cooperación entre bibliotecas u otro tipo de instituciones.

2.3.1 International Internet Preservation Consortium

La organización que agrupa a la mayor parte de estas iniciativas es el IIPC (International Internet Preservation Consortium, http://netpreserve.org), que tiene la misión de adquirir, preservar y hacer accesible el conocimiento y la información sobre Internet para las futuras generaciones de todo el mundo, promoviendo el intercambio global y las relaciones internacionales.

Fue creado formalmente en julio de 2003 por los 12 miembros que actualmente forman el consorcio: Bibliothèque Nationale de France (http://www.bnf.fr) (coordinador), Biblioteca Nazionale Centrale di Firenze (http://www.bncf.firenze.sbn.it), Det Kongelige Bibliotek (http://www.kb.dk), Helsingin yliopiston kirjasto-Suomen Kansalliskirjasto (http://www.lib.helsinki.fi), Internet Archive (http://www.archive.org), Kungliga biblioteket Sveriges nationalbibliotek (http://www.kb.se), Landsbokasafn Islands-Haskolabokasafn (http://www.bok.hi.is), Library and Archives Canada (http://www.collectionscanada.ca), Nasjonalbiblioteket (http://www.nb.no), National Library of Australia (http://www.nla.gov.au), The British Library (http://www.bl.uk) y The Library of Congress (http://www.loc.gov).

El IIPC tiene los siguientes objetivos:

- Recoger una parte rica del contenido de Internet de todo el mundo, para ser preservada de forma que pueda ser archivada, preservada y asegurado el acceso en el tiempo.
- Fomentar el desarrollo y el uso de las herramientas comunes, técnicas y estándares que permitan la creación de archivos internacionales.
- Animar y ayudar a las bibliotecas nacionales de todo el mundo para archivar y preservar Internet.

Existen diversos grupos de trabajo (herramientas de acceso, gestión de contenidos, etc.) creados al amparo del consorcio, y con la intención de publicar informes y facilitar el acceso a software, cuestiones éstas que no han sido públicamente completadas.

Así pues, el consorcio no captura webs, sino que agrupa a una serie de instituciones que lo hacen, y tiene como objetivo promover estas actividades.

2.3.2 Nordic Web Archive

El NWA (Nordic Web Archive, http://nwa.nb.no) es un foro de les bibliotecas nacionales escandinavas (Dinamarca, Finlandia, Islandia, Noruega y Suecia) para la coordinación y el intercambio de experiencias en los campos de la captura y el almacenaje de documentos web.[15]

Desde noviembre de 2000 se ha desarrollado el conjunto de herramientas NWA:[16] un software para acceder a los documento web archivados, creado usando PHP, Perl y Java, con estándares abiertos como el protocolo HTTP y XML para la comunicación entre las diferentes partes del sistema. El uso del paquete de software (búsqueda y navegación por el archivo web), se realiza por medio de un buscador web estándar, y no es necesario ningún *plugin* específico.

La iniciativa se debe a Nordunet2 (programa de investigación de los escandinavos), Nordinfo (consejo escandinavo para la información científica que incluye a las bibliotecas de investigación) y las bibliotecas nacionales escandinavas.

2.3.3 Internet Archive

El Internet Archive (http://www.archive.org) es una organización sin ánimo de lucro fundada en 1996 para construir una "biblioteca de Internet" y ofrecer acceso permanente a investigadores, historiadores, personal académico y al público en general a las colecciones históricas en formato digital. Situado en la antigua prisión de San Francisco (EEUU), el archivo ha recibido donaciones de IBM, Alexa (filial de Amazon) y otras organizaciones que han facilitado su crecimiento.

Recibe el apoyo de diversos organismos como la Library of Congress, los US National Archives y los UK National Archives, entre otros.

Actualmente Internet Archive se considera el mayor archivo web del mundo,[17] e incluye texto, audio, imagen en movimiento y software, así como páginas web archivadas de todo el mundo, inclusive un número representativo de recursos catalanes.[18] De acceso abierto y en línea, el gigante contiene en un petabyte un total aproximado de 600 millones de sedes web, desde 1996 hasta la actualidad, y cada dos meses se realiza una captura masiva que afecta a millones de páginas web (crecimiento mensual de 20 Tbytes), siguiendo el modelo exhaustivo que Suecia y otros países representan.

El programa Heritrix (http://crawler.archive.org) es el gestor (software libre) que utiliza el Internet Archive, y el sistema de depósito se realiza en múltiples copias, separadas geográficamente.

Recientemente y con sede en Ámsterdam se ha creado el European Digital Archive, rama europea del Internet Archive.

3 El futuro es generalizado, híbrido, costoso y cooperativo: conclusiones sobre la panorámica

El interés por la preservación digital está ya **generalizado** en los países desarrollados, pese a que con un grado de desarrollo heterogéneo. Posiblemente en el momento de publicación del presente artículo existan más proyectos que los veinte mencionados, aunque algunos estén en fase de diseño.

El futuro es **híbrido**: la diferenciación en modelos (integral *versus* selectivo) representa sólo la primera fase de desarrollo de los proyectos.

Los proyectos de depósito son económicamente **costosos** y pasan forzosamente por la implicación del número más elevado de agentes posibles que doten de continuidad a los programas una vez iniciados. En este sentido, los fracasos que regularmente afectan a los proyectos estudiados se deben a la falta de financiación.

Existe una corriente global de **cooperación** (compartir experiencias, el relato de los éxitos y fracasos, software de código abierto) entre los proyectos. Sirva como ejemplo más evidente la generalización del software Heritrix.

Los **acuerdos** con los productores y editores web son garantía de éxito. No siempre una ley moderna de depósito legal acompaña a los depósitos digitales nacionales que existen, y el acceso a la infranet (y a la Internet invisible) ha de contemplarse, con o sin ley.

En Cataluña, la Biblioteca de Catalunya ha iniciado con su proyecto PADICAT (Patrimoni Digital de Catalunya) las actuaciones necesarias para hacer siempre accesibles las webs catalanas.

Fecha de recepción: 03/10/2005. Fecha de aceptación: 25/10/2005.

Anexo. Descripción sumaria de los proyectos de depósitos nacionales

AOLA (http://www.ifs.tuwien.ac.at/~aola)

- Liderado por la Österreichische National-bibliothek (Austria).
- Iniciado en 1999.
- Modelo integral con elementos del modelo híbrido.
- El crecimiento previsto es de 7 Gb diarios.
- El proyecto ha sufrido paradas por falta de fondos.
- El software NEDLIB de la primera fase dejó lugar al Combine en una etapa posterior.

[*Archive of Czech web resources*] (http://webarchiv.nkp.cz/index-e.html)

- Liderado por la Národní knihovna Ceské Republiky (República Checa).
- Iniciado en 2000.
- Modelo integral.
- En colaboración con otras instituciones bibliotecarias y de investigación, la biblioteca nacional checa ha impulsado las capturas anuales del dominio .cz por medio de una adaptación del software NEDLIB.
- Está previsto incluir publicaciones digitales en futuras etapas.

[*Archiving the French web*] (http://www.bnf.fr)

- Liderado por la Bibliothèque Nationale de France (Francia).
- Iniciado en el año 2000.
- Modelo híbrido.
- El alcance del proyecto incluye capturas automáticas a gran escala, capturas sistemáticas y continuas de una selección de sedes web (el 10 % del total), depósito de la Infranet, y capturas temáticas de sedes web muy efímeras (elecciones francesas de 2002: 1.900 sedes web).

[*Archiving the Web Greek*]

- Liderado por la Athens University of Economics and Business (Grecia).
- Iniciado en 2003.
- Modelo integral.
- El proyecto griego es un experimento destinado a capturar el dominio griego, con software propio.

Deposit.ddb.de (http://deposit.ddb.de/online/vdr/titel.htm)

- Liderado por Die Deutsche Bibliothek (Alemania).
- Iniciado en 1997.
- Modelo integral con elementos del modelo híbrido.
- Catalogación por metadatos.
- Las pruebas iniciales se realizaron con la web del Gobierno alemán.
- A partir de 2002 se llega a acuerdos con editores alemanes.

E-Collection (http://epe.lac-bac.gc.ca/)

- Liderado por Libraries and Archives Canada (Canadá).
- Iniciado en 1994.
- Modelo selectivo.
- A partir del EPPP (*Electronic Publications Pilot Project*), de 1994–95, se creó la E-Collection, destinada al archivo en línea de publicaciones digitales, a texto completo.
- La actualización del proyecto, 2004-05, incluye tesis en línea, webs, etc.

e-Depot (http://www.kb.nl/dnp/e-depot/e-depot-en.html)

- Liderado por la Koninklijike Bibliotheek (Países Bajos).
- Iniciado en 1995.
- Modelo selectivo.
- A partir de los acuerdos con los editores (y en los Países Bajos se ubican un número importante de multinacionales editoras), se apunta a les revistas publicadas en Holanda, donde no existe legislación estricta de depósito legal.
- Contiene tres millones de números de revistas (marzo de 2005).

EVA (http://www.lib.helsinki.fi/tietolinja/0203/webarchive.html)

- Liderado por la Helsingin yliopiston kirjasto (Finlandia).
- Iniciado en 1997.
- Modelo integral con elementos del modelo híbrido.
- La biblioteca nacional de Finlandia lidera el NWA (Nordic Web Archive), que pretende ser el archivo web escandinavo.
- El proyecto EVA dirigido en sucesivas etapas a publicaciones periódicas, dio paso en 2001 al archivo web que incluye el dominio .fi.

IRIS (http://catalogue.bnquebec.ca:4400/cap_fr.html)

- Liderado por la Bibliothèque nationale du Québec (Québec).
- Iniciado en 2000.
- Modelo selectivo.
- 3.469 monografías y 1.100 títulos de revistas digitales (octubre de 2004) forman el cuerpo del archivo, que está integrado en el catálogo IRIS.
- Captura con acuerdos con el gobierno quebequés en una primera fase.

LIBIS Electronic Resources Subsystem (http://www.libis.lt/en/welcome.html)

- Liderado por la Martynas Mazvydas (Lituania).
- Iniciado en 2002.
- Modelo integral.
- El proyecto LIBIS consiste en completar el catálogo LIBIS con las capturas procedentes del sistema NEDLIB.
- El proyecto incluye metadatos Dublin Core.

Minerva (http://www.loc.gov/minerva/)

- Liderado por la Library of Congress (Estados Unidos de América).
- Iniciado en 2000.
- Modelo selectivo temático.
- Asociado al gigante Internet Archive, el recurso Minerva captura selectivamente 35 sedes web.
- En las elecciones presidenciales de 2000 (y en otras fechas especiales, como el 11S), se aumenta la captura selectiva.
- Está previsto incluir publicaciones digitales en futuras etapas.

New Zealand's digital Heritage (http://www.natlib.govt.nz/bin/media/pr?item=1085885702)

- Liderado por la National Library of New Zealand (Nueva Zelanda).
- Iniciado en 1999.
- Modelo híbrido.
- La ley de depósito legal neozelandesa, de 2003, abría el panorama a los recursos en línea, incluidas las publicaciones en abierto y también la infranet.
- Presupuesto global patrimonio digital (2004): 14 M€.
- Usa el software Pandas, pero el acceso no es en abierto.

Paradigma (http://www.nb.no/paradigma/eng_index.html)

- Liderado por la Nasjonalbiblioteket (Noruega).
- Iniciado en 2001.
- Modelo integral con elementos del modelo híbrido.
- Con capturas anuales, a partir de 2003 se incluyó la captura de los dominios internacionales con contenido noruego, así como 65 diarios digitales.
- La indización de los recursos se produce automáticamente por el software FAST.

UK Web Archive (http://www.webarchive.org.uk/)

- Liderado por la British Library (Reino Unido).
- Iniciado en 2004.
- Modelo selectivo.
- Siguiendo el modelo australiano, con el mismo software Pandas y una interfaz de consulta muy similar, en mayo de 2005 presentaba 1.030 sedes web con las que se ha llegado a acuerdos.
- En la selección, y en todo el proyecto, participan el resto de bibliotecas nacionales del Reino Unido.

WARP (http://warp.ndl.go.jp)

- Liderado por la National Diet Library (Japón).
- Iniciado en 2002.
- Modelo selectivo.
- Las enmiendas a la ley de depósito legal de 2000 incluyen CDR y otros materiales digitales en soportes físicos.
- La biblioteca nacional del Japón recoge (junio de 2004) con acuerdos 600 sedes web (administración, universidades, empresas, etc.), y 110 diarios electrónicos.

Notas

[1] *Directrices para la preservación del patrimonio digital* (Canberra: Unesco, 2003), <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 03/10/2005].

[2] El UK Web Archiving Consortium fija en 44 días la media de vida de una página web <http://info.webarchive.org.uk/pressrelease21-06-04.html>. [Consulta: 03/10/2005].

[3] El año 1999 la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona tuvo una iniciativa en esta misma línea al organizar un seminario sobre esta cuestión y publicar: *Biblioteques digitals i dipòsits nacionals de recursos digitals* (Barcelona: Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, 1999).

[4] *Dipòsit* es la palabra catalana normalizada que designa al *repository* en inglés.

[5] José Antonio Cordón, "El depósito legal y los recursos digitales en línea". En: *Las bibliotecas nacionales del siglo XXI* (Valencia: Biblioteca Valenciana, 2005), <http://bv.gva.es/documentos/Ponencias/Cordon.pdf>. [Consulta: 03/10/2005].

[6] La traducción de *nu* es *ahora*. Pese a que sea el dominio geográfico de la isla Nieu, en Polinesia, en sueco y otras lenguas escandinavas es muy utilizada para dotar de un componente dinámico el nombre del dominio.

[7] Útiles como Maxmind (http://www.maxmind.com) o Ip2location (http://www.ip2location.com) facilitan la ubicación geográfica de los servidores.

[8] InternetLab (http://internetlab.cindoc.csic.es) pertenece al CINDOC-CSIC.

[9] *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[10] Un recurso que proporciona información detallada de los proyectos nacionales es *PADI: Preserving Access to Digital Information* (http://www.nla.gov.au/padi/) de la National Library of Australia. PADI contiene información actualizada de la práctica totalidad de los proyectos existentes, así como información

diversa de los aspectos relacionados con la preservación web (depósito legal, herramientas, bibliografía, etc.). También la National Library of Australia organizó en noviembre de 2004 el congreso *Archiving web resources*, en el cual es accesible en abierto la mayor parte de presentaciones realizadas en aquella actividad. Finalmente, también cabe tener presente el IWAW (International Web Archiving Workshop, http://www.iwaw.net) que se celebra anualmente organizado por un grupo de profesionales procedentes de los diversos proyectos. En este espacio web también se puede consultar la documentación relativa a los proyectos existentes. El crecimiento del número de proyectos destinados a crear depósitos nacionales es constante, y próximamente se habrán de sumar, a los expuestos en este artículo los de las bibliotecas nacionales de Luxemburgo, Singapur, Egipto, Croacia, y evidentemente, el de la Biblioteca de Catalunya.

[11] Joan Mannerheim, "Collect all, catalogue some". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[12] Margaret Phillips, "What to collect and how to do it: the National Library of Australia's selective approach". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[13] La fecha tiene carga simbólica: la fiesta de San Juan (el solsticio de verano) es especialmente celebrada en los países escandinavos.

[14] "New legal deposit law", *Netarchive.dk*, <http://netarchive.dk/newsite/news/index-en.php>. [Consulta: 03/10/2005].

[15] Porsteinn Hallgrímsson, Sverre Bang, "Nordic Web Archive". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[16] El software se puede descargar en Sourceforge (http://nwatoolset.sourceforge.net).

[17] Michele Kimpton, "Saving the web for future generations". En: *Archiving web resources* (Canberra: National Library of Australia, 2004), <http://www.nla.gov.au/webarchiving/>. [Consulta: 03/10/2005].

[18] Algunos ejemplos (agosto 2005) son las páginas web de la Generalitat de Catalunya (http://www.gencat.net), con 207 capturas desde mayo de 2002; el diario *Avui* (http://www.avui.es), con 55 capturas desde enero de 1998; la Universitat de Barcelona (http://www.ub.es y http://www.ub.edu), con 223 capturas desde febrero de 1997, etc.