El *International Internet Preservation Consortium* (IIPC) y el papel de la Biblioteca Nacional en el archivo de Internet

BIBLIOTECA NACIONAL

BN

Teresa Malo de Molina
Directora Técnica

¿Qué es el IIPC?

Misión:

- Adquirir, preservar y hacer accesible la información y el conocimiento que hay en Internet para las futuras generaciones de todo el mundo, promoviendo el intercambio global y las relaciones internacionales

¿Qué es el IIPC?

Objetivos principales:

- Preservar la colección de un cuerpo importante de contenidos de Internet de todo el mundo de manera que puedan ser almacenados de forma segura y accesible en el futuro
- Promocionar el desarrollo y el uso de herramientas, técnicas y estándares que capaciten la formación de archivos internacionales
- Animar y apoyar a las bibliotecas nacionales de todo el mundo a dirigir la preservación y el archivo de Internet

¿Qué es el IIPC?

Otros objetivos:

- Proporcionar un foro para compartir conocimientos acerca del archivo de Internet entre los miembros y también con otras instituciones
- Desarrollar y recomendar estándares
- Desarrollar herramientas ínteroperativas y técnicas para adquirir, archivar y proporcionar acceso a sitios web
- Crear conciencia de la importancia de la preservación de los contenidos de Internet en Conferencias, Seminarios, Publicaciones, etc.

¿Qué es el IIPC?

- Coodinador: Bibliothèque Nationale de France
- Participantes:
  - Biblioteca Nazionale Centrale di Firenze
  - Det Kongelige Bibliotek (Real Biblioteca de Dinamarca)
  - Helsingin yliopiston kirjasto-Suomen Kansalliskirjasto (Biblioteca Universitaria de Helsinky-Biblioteca Nacional de Finlandia)
  - Internet Archive
  - Kungliga biblioteket Sveriges nationalbibliotek (Biblioteca Nacional de Suecia)
  - Landsbokasafn Islands - Haskolabokasafn (Biblioteca Nacional de Islandia)
  - Library and Archives Canada
  - Nasjonalbiblioteket (Biblioteca Nacional de Noruega)
  - National Library of Australia
  - The British Library
  - The Library of Congress

¿Qué es el IIPC?

- El IIPC se crea en julio del 2003 con 12 instituciones participantes
- El acuerdo inicial se firma por 3 años y se decide no ampliar los miembros en este periodo
- A partir de julio de 2006 se abre la participación a nuevos miembros
- La Biblioteca Nacional de España ha pedido ya su incorporación a este Consorcio

¿Qué es el IIPC?

- Hay dos niveles de trabajo:
  - A través de Grupos de Trabajo específicos
  - A través de proyectos aceptados por el Comité de Dirección
- Se espera obtener:
  - Herramientas desarrolladas bajo licencias gratuitas de código abierto
  - Recomendaciones (metodologías, procesos, estándares)

Grupos de Trabajo

- **Estructura**: arquitectura y estándares
- **Indicadores y plataformas de experimentación**: definición e implementación de un entorno de experimentación para los metabuscadores (crawlers)
- **Herramientas de acceso**: desarrollo de un conjunto de herramientas

Grupos de Trabajo

- **Web profundo**: desarrollo de herramientas para depositar y acceder a plataformas documentales construidas sobre bases de datos
- **Gestión de contenido**: visión común de la cobertura de colecciones y la complementariedad. Estadísticas
- **Necesidades de los investigadores**: comentarios y consejos sobre contenido y acceso

Avances en la Estandarización

- Módulos funcionales y arquitectura con estándar de APIs
  - Para permitir la interoperabilidad con el sistema de cada institución
  - Con una estructura modular para permitir el desarrollo de nuevas capacidades
- Formato para el archivo web y el intercambio
  - Desde el formato ARC al WARC, intentando que se introduzca como norma ISO TC 46/SC4
- Metadatos para la conservación a largo plazo
- Identificación permanente

Resultados

- Descarga de programas:
  - http://netpreserve.org/software/downloads.php
- El IIPC Toolkit estará listo a finales del 2006:
  - Robusto y escalable para todo el web
  - Implementando estándares IIPC (ARC 3.0, Metadatos, API…)
  - Fácil de instalar y usar por usuarios avanzados (ingenieros de archivos de Internet
  - En código abierto y disponible para la comunidad de los archivos web

Otros resultados

- WERA
- WARC
- Nuevo conjunto de Metadados para el Archivo de Internet
- NUTCH

¿Qué es WERA?

http://nwa.nb.no/

- Una herramienta de acceso a colecciones de archivo web, lo que se conoce como un visor WAC (Web Archive Collection),
- Como Internet Archive Wayback Machine pero…
  - Soporta búsqueda a texto completo en las colecciones de archivo web
  - Es código libre, se puede distribuir
- Su desarrollo está patrocinado por el IIPC
- Es un subconjunto del conjunto de herramientas NWA (Nordic Web Archive)
- Primera versión en agosto de 2005

¿Qué es WARC?

http://cvs.sourceforge.net/viewcvs.py/archive-access/archive-access/src/docs/warc/

- WARC = Web ARChive file format
- Una nueva generación de ARC, cuyo nombre ha creado el IIPC
  - El formato ARC fue creado por el Internet Archive
  - Se han creado más de 600TB de ficheros ARCs desde 1996
- Un fichero WARC o ARC es una secuencia simple de bloques de contenido, cada bloque va precedido por una pequeña cabecera textual
  - ARC es para los que los crawlers capturen contenidos fácilmente
  - WARC es para capturar y relacionar bloques de contenido

¿Qué es NUTCH?

http://lucene.apache.org/nutch/

- Un proyecto reciente de código abierto
- Una aplicación para la búsqueda en Web
- Un pequeño pero creciente grupo de usuarios y desarrolladores
- Detrás de unos pocos sitios web
- Un proyecto Apache
- Construido sobre Lucene
- No es:
  - Un negocio
  - Un buscador web, pero quiere potenciar muchos buscadores web, desde dominios específicos hasta el web global
  - Un proyecto de investigación, pero quiere ser una plataforma de investigación
- En el futuro será NDFS (Nutch Distributed File System)

El futuro

- Los tres primeros años del Consorcio se han dedicado a la creación de un conjunto de herramientas, en paralelo con una gran actividad de estandarización
- En la siguiente fase se construirá la primera capa de herramientas en la que se incluirán las más sofisticadas dedicadas a la adquisición y el acceso
- El trabajo sobre preservación digital ya iniciado también será una pieza clave en las actividades futuras del Consorcio

¿Qué está haciendo la Biblioteca Nacional de España?

- Implementar la infraestructura tecnológica necesaria
- Incorporarse al IIPC y colaborar en todas las iniciativas relacionadas (el proyecto de Biblioteca Digital Europea)
- Comenzar con una experiencia piloto selectiva (tema o evento)
- Intentar incorporar este concepto en el nuevo reglamento del depósito legal que se va a desarrollar a partir de la aprobación de la nueva Ley del Libro y Promoción de la Lectura

Conclusiones

- El archivo de Internet está todavía en su más temprano inicio y solo una pequeñísima parte del web global está siendo recogido y preservado. La institución pionera es el Archivo de San Francisco, quien cuenta con el mayor archivo de sitios webs recolectados de todas las partes del mundo. A pesar de ello, es solo una mínima parte de lo que está disponible.
- Es evidente que es prácticamente imposible que una ley obligue a que todo aquel que publique cualquier contenido en Internet deba enviar una copia a la Biblioteca depositaria del Depósito Legal

Conclusiones

- La única solución práctica posible es usar los métodos automáticos de recolección para recoger documentos webs, y que las leyes de depósito legal en cada país permitan hacer esto a las bibliotecas nacionales. Sin embargo, esto tardará mucho tiempo en ser efectivo, incluso en aquellos países donde la biblioteca nacional pueda decir algo respecto a esta ley.
- Es muy difícil establecer fronteras nacionales en la web y por ello el IIPC está trabajando en definir y desarrollar los necesarios componentes tecnológicos y los procedimientos y estándares que permitirán la construcción de archivos nacionales y globales con el acceso necesario

Conclusiones

- Se utilizan dos métodos de recolección:
  - Por una parte, una recolección exhaustiva que pretende obtener una colección representativa del material publicado en el web
  - Por otra, una recolección selectiva que recoge en profundo para obtener una idea precisa de lo que contiene un determinado sitio web en el momento en que se recolecta.
- Ambos métodos se combinan en las tres políticas de colección actualmente utilizadas:
  - Cuando se pretende proporcionar una fotografía instantánea de un determinado dominio web, se utiliza una recolección exhaustiva
  - Cuando se quiere recopilar un alto porcentaje del contenido de un número limitado de sitios webs escogidos, se utiliza una recolección selectiva
  - Cuando se intenta recoger a fondo contenidos dedicados a un tema o evento específico, se utiliza una recolección combinada

Conclusiones

- La tareas relacionadas con el Archivo de Internet se mueven en las fronteras de dos profesiones distintas: los bibliotecarios y los especialistas en tecnologías de la información, y los diferentes métodos existentes reflejan esta condición porque es una colección bibliotecaria pero requiere una implicación sustancial de las TIC:
  - Así por un lado existen experiencias como el proyecto australiano **Pandora** con valores bibliotecarios tradicionales donde se seleccionan y catalogan sitios web de *calidad* y el acceso es a través de una búsqueda estructurada.
  - Y por otro lado, está el **Internet Archive** que usa un sistema de recolección cruzada de material web publicado en todo el mundo, sin tener en cuenta la legislación de los diferentes países, y el acceso se realiza a través de sistemas automáticos de indización

Conclusiones

- El acceso a los archivos del web (la navegación, la indización y las búsquedas) no es autoevidente y las reglas para el acceso general difieren mucho en los distintos países, pero en la mayoría de los casos está orientado al uso de los investigadores. Esto tendrá que cambiar en el futuro si se quiere que los archivos de Internet tengan un uso óptimo

Conclusiones

- Los archivos web contienen documentos multimedia y de momento sólo es posible indizar texto. Para otro tipo de documentos, como las imágenes, la indización exige recolectar metadatos con información textual añadida en el fichero de cabecera. Sin embargo, se está investigando mucho en la indización automática de sonidos e imágenes fijas y en movimiento, y cuando estos métodos sean una realidad, se podrán indizar estas partes de los archivos web.

Conclusiones

- Aparte del trabajo en marcha, es importante señalar que desde el punto de vista de la recolección y la preservación de Internet nuestra comprensión del medio y sus contenidos deja mucho que desear: todas las herramientas actuales necesitan ser mejoradas y hay que desarrollar nuevas herramientas.
- La labor del IIPC servirá para aumentar la concienciación y la actividad en el archivo del web, y unirá los avances tecnológicos con la cooperación internacional, lo que permitirá, en un futuro cercano, que las bibliotecas nacionales del mundo recopilen y preserven los contenidos de Internet igual que hacen hoy con las colecciones impresas.
- Las bibliotecas nacionales deberán considerar el Archivo de Internet y su indización, a pesar del enorme cambio que supone, como una gran oportunidad, en lugar de contemplarlo como un problema o un incordio

Direcciones de interés

- International Internet Preservation Consortium (IIPC) (http://www.netpreserve.org/)
- Bibliothèque Nationale de France (http://www.bnf.fr/)
- Biblioteca Nacional de España (http://www.bne.es)
- Biblioteca Nazionale Centrale di Firenze (http://www.bncf.firenze.sbn.it/)
- Det Kongelige Bibliotek (Real Biblioteca de Dinamarca) (http://www.kb.dk/)
- Helsingin yliopiston kirjasto-Suomen Kansalliskirjasto (Biblioteca Universitaria de Helsinky-Biblioteca Nacional de Finlandia) (http://www.lib.helsinki.fi/)
- Internet Archive (http://www.archive.org/)

Direcciones de interés

- Kungliga biblioteket Sveriges nationalbibliotek (Biblioteca Nacional de Suecia) (http://www.kb.se/)
- Landsbokasafn Islands - Haskolabokasafn (Biblioteca Nacional de Islandia) (http://www.bok.hi.is/)
- Library and Archives Canada (http://www.collectionscanada.ca/)
- Nasjonalbiblioteket (Biblioteca Nacional de Noruega) (http://www.nb.no/)
- National Library of Australia (http://www.nla.gov.au/)
- The British Library (http://www.bl.uk/)
- The Library of Congress (http://www.loc.gov/)
- ARC File Format (http://www.archive.org/web/researcher/ArcFileFormat.php)

Artículos de interés

- Halgrimsson, Thorsteinn: Special PresentationThe International Internet Preservation Consortium (IIPC) (http://consorcio.bn.br/cdnl/2005/HTML/Presentation%20Thorsteinn %20Halgrimsson.htm)
- Henriksen, Birgit N.: Webarkivering (http://netarkivet.dk/website/publications/ webarkivering-webarchiving.pdf)
- Lupovici, Catherine: Web archives long term access and interoperability: the International Internet Preservation Consortium activity (http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf)

¡Muchas gracias!

directora.tecnica@bne.es