netarchive.dk

La Captura de Internet en Dinamarca – los primeros dos años

El 1 de julio de 2005, entró en vigor la última ley danesa de Depósito Legal. El Capítulo 3 de la ley se refiere a los “Materiales publicados en la red de comunicaciones electrónica” y permite la captura de Webs dentro del dominio <.dk> y Webs de otros dominios que se dirijan al público danés. Desde esa fecha, la Royal Library (Biblioteca Nacional de Dinamarca) y la State and University Library han capturado en Internet más de 34 terabytes. En este artículo se abordan las experiencias acumuladas por las bibliotecas a lo largo de los dos primeros años.

Estrategias de captura

Antes de que la ley se aprobara y entrara en vigor, algunos experimentos habían puesto de manifiesto que, para capturar la Web, debían aplicarse tres tipos de estrategias: una foto fija general de todo el dominio <.dk> hecha cuatro veces al año, una captura selectiva pero más frecuente de Webs dinámicas y una captura irregular de Webs seleccionadas relacionados con determinados acontecimientos.

Captura de foto fija

La idea que subyace tras la foto fija consiste en obtener una panorámica completa del dominio danés capturando todas las páginas Web cuatro veces al año. Hasta la fecha, hemos podido completar tres capturas, una en 2005, dos en 2006, mientras que en marzo de 2007 se inició una cuarta captura. Con la capacidad de hardware disponible, ahora podemos realizar foto fija completa en un plazo de tres meses. De este modo, a partir de 2008, deberíamos poder hacer cuatro fotos fijas al año. Sigue abierta la cuestión de si podemos reducir este lapso de tiempo, y en qué medida, para obtener una foto fija completa, ya que también es necesario tener en cuenta el tráfico en las Webs que capturamos. Nuestro deseo es no tener que cerrar ninguna Web mientras realizamos la captura, por lo que, especialmente en el caso de Webs más pequeñas, intentamos hacer la recopilación a una velocidad que no obstaculice el tráfico normal en Internet.

Después de las dos primeras capturas de foto fija (otoño de 2005 y primavera de 2006), nos centramos en la integración en nuestro sistema de un módulo de desduplicación. Esta tarea se completó con éxito a lo largo del verano de 2006 y ha generado una reducción significativa (31%) de los bytes que deben archivarse como resultado de una captura de foto fija. Este porcentaje aumentará cuando alcancemos el objetivo de realizar cuatro fotos fijas al año, aumentando así la frecuencia de duplicaciones.

Antes de iniciar una captura, el sistema debe disponer de un listado de dominios que serán objeto de dicha captura. Este listado es facilitado por el host master del dominio <.dk> que está obligado, con

arreglo a la ley, a facilitar a netarchive un listado de todos los dominios registrados dentro del dominio de nivel superior (Top Level Domain) <.dk>. Este método ha funcionado muy bien, y hemos podido volcar una lista actualizada en el sistema antes de cada captura. La primera captura se hizo por fases, con el fin de desvelar *crawler traps* o trampas para la araña (como por ejemplo calendarios). Durante la primera fase, se recopilaron todas las Webs con un máximo de 10 objetos y comprobamos que aproximadamente 235.000 Web entraban dentro de este límite. Durante la siguiente fase, se estableció un límite de 50 objetos por Web, posteriormente de 500 objetos y finalmente de 5.000 objetos. Durante estas capturas, se capturaron más del 99% de los dominios daneses en su totalidad. Dos miembros del personal examinaron manualmente las Webs que contenían más de 5.000 objetos, un total de 3.500, para detectar *crawler traps*, alias y otros problemas, que fueron posteriormente eliminados (URLs excluidos de la captura) y se recopilaron las Webs en su totalidad.

Tras la primera captura de foto fija, decidimos cambiar el parámetro de medición de límites, pasando de objetos a Mbytes, de manera que la primera ronda de captura tenía un límite de 10 Mbytes. Más de 600.000 Webs estaban por debajo de este límite y unas 61.000 llegaban al límite. El límite para la segunda ronda se estableció en 500 Mbytes, por lo que quedaban unas 6.300 Webs para ser examinadas manualmente. Esto se hizo para eliminar los problemas más arriba mencionados, pero también para determinar si determinadas Webs, por lo general páginas particulares con fotos y videos de reuniones familiares, bodas, etc., podían ser objeto de un muestreo en lugar de ser recopilados en su totalidad. De momento, decidimos limitar esas Webs a 499 MB. Esto redujo a 263 Webs el número de Webs que debían capturarse hasta un límite de 1 Gbyte. La Web más grande, con mucha diferencia, es la de la corporación nacional de radiotelevisión, con más de 150 GB (www.dr.dk).

La cuarta foto fija se ha iniciado sin proceder por pasos (primero hasta 10MB – y posteriormente sólo dominios que llegaban al límite de 10MB). Sin embargo, en la práctica no ha resultado una buena idea (quedando así demostrado que la captura por fases es una buena solución), porque las Webs grandes y las pequeñas se mezclan en la misma operación de captura, lo que resta eficacia al trabajo. Por tanto, a partir de la próxima foto fija retomaremos la estrategia original, cubriendo todas las Webs pequeñas en una primera aproximación.

Captura selectiva

El objetivo de la captura selectiva es reunir páginas Web que se actualizan frecuentemente y que no quedarían reflejadas en la captura de foto fija.

Estos tipos se han definido como sigue:

- Sitios nuevos (medios de comunicación nacionales y regionales)
- Webs “típicas” dinámicas y muy visitadas que representan a la sociedad civil, al sector comercial y a la administración pública.
- Sitios experimentales y/o únicos, que documentan nuevas formas de utilizar la Web (por ejemplo, arte en la Web)

La política actual consiste en recopilar unas 80 Webs de este tipo. Seleccionar tan sólo 80 Webs entre todas las Webs danesas no es una tarea fácil, pero está cifra se ha determinado teniendo en cuenta los recursos disponibles. Las Webs nuevas son la categoría más fácil de seleccionar, y de momento son unas 30-35 Webs (Dinamarca es un país pequeño). Sin embargo, el mercado de los medios de comunicación cambia constantemente y el desarrollo debe ser objeto de un seguimiento continuo. Las otras dos categorías son más difíciles de seleccionar. Todos sabemos que la Web ha impulsado nuevas formas de comunicación humana (comunidades virtuales de todo tipo) y es importante documentar esto. Un Consejo Asesor Editorial, formado por investigadores y profesionales de los medios de comunicación, colabora con netarchive en el proceso de selección (ver más abajo).

Una vez se ha seleccionado una Web, ésta será objeto de un examen minucioso para determinar qué partes de la Web deben ser capturadas y con qué frecuencia. La frecuencia puede varias de varias veces al día a una vez a la semana. No resulta sorprendente que este tipo de captura requiera un estrecho seguimiento por parte del personal para encontrar las Webs y determinar la frecuencia y la profundidad de la captura (las Webs nuevas, por ejemplo, pueden tener bases de datos con noticias y artículos antiguos, que no cambian, y se recopilarán mediante la captura de foto fija, por lo que únicamente las “portadas” de la Web deben ser capturadas frecuentemente). La reducción de duplicación ha tenido un verdadero impacto en este tipo de capturas, registrándose un ahorro del 50-70% de bits a archivar, lo que también supone que podemos hacer una captura más en profundidad de lo estrictamente necesario para cubrirnos las espaldas.

Captura de acontecimientos

El objetivo de esta estrategia es poder recopilar páginas Web a partir de nuevas Webs dedicadas a un acontecimiento y que desaparecerán una vez finalizado el acontecimiento.

Hemos definido un acontecimiento como algo que

- Crea un debate entre la población y se prevé que sea relevante para la historia danesa o que tenga un impacto en el desarrollo de la sociedad danesa
- Genere la aparición de nuevas Webs dedicadas al acontecimiento
- Se aborda de manera exhaustiva en otras Webs existentes

El primer acontecimiento que capturamos fueron las elecciones locales de noviembre de 2005. Al margen de tratarse de unas elecciones, que siempre son de interés para los especialistas, en este caso tenían una relevancia especial debido a la decisión del Gobierno de reestructurar la Administración local del país, reduciendo el número de municipios (*kommuner*) de 271 a 98, y sustituyendo 14 condados (*amter*) por 5 regiones. Esto supone también un menor número de alcaldías y concejalías para los políticos locales, así como nuevas constelaciones políticas de votantes en un municipio. En relación con unas elecciones previas capturadas como proyecto piloto en 2001, las elecciones de 2005 pusieron de manifiesto un importante aumento de las Webs de candidatos individuales. Más de 1.000 candidatos disponían de su propia Web, la mayoría de las cuales desaparecieron después de las

elecciones, y se espera que este modelo de comportamiento se convierta en la norma en futuras elecciones.

El segundo acontecimiento que optamos por capturar fue la crisis desatada en el invierno 2005/2006 por la publicación de caricaturas de Mahoma en un periódico danés.

El tercer acontecimiento consistió en un seguimiento de las elecciones de 2005. La nueva estructura local entró en vigor el 1 de enero de 2007 y durante los últimos meses de 2006 hicimos una captura especial de todas las Webs municipales y de los condados ya que preveíamos -con razón- que las Webs de los antiguos municipios desaparecerían del mismo modo que las Webs creadas para informar a los ciudadanos de los cambios en la administración local y regional.
En la actualidad estamos recopilando acontecimientos menos importantes que generan la aparición de nuevas Webs durante un período breve (y luego desaparecen) entre fotos fijas. Un ejemplo: el desalojo, el 1 de marzo de 2007, por la policía de una casa en Copenhague ocupada por un grupo de jóvenes generó una serie de páginas Web relacionadas con este acontecimiento.

Cabe subrayar que todo el material capturado se fusionará en un único archivo, con independencia de cómo ha sido capturado. Las estrategias expuestas son estrategias para recopilar el material, no para desarrollar colecciones. No obstante, sí mantenemos un registro de nuestra actividad de captura como documentación para futuros usuarios sobre cómo se recopiló el contenido del archivo.

Materiales capturados – estadísticas y tipos

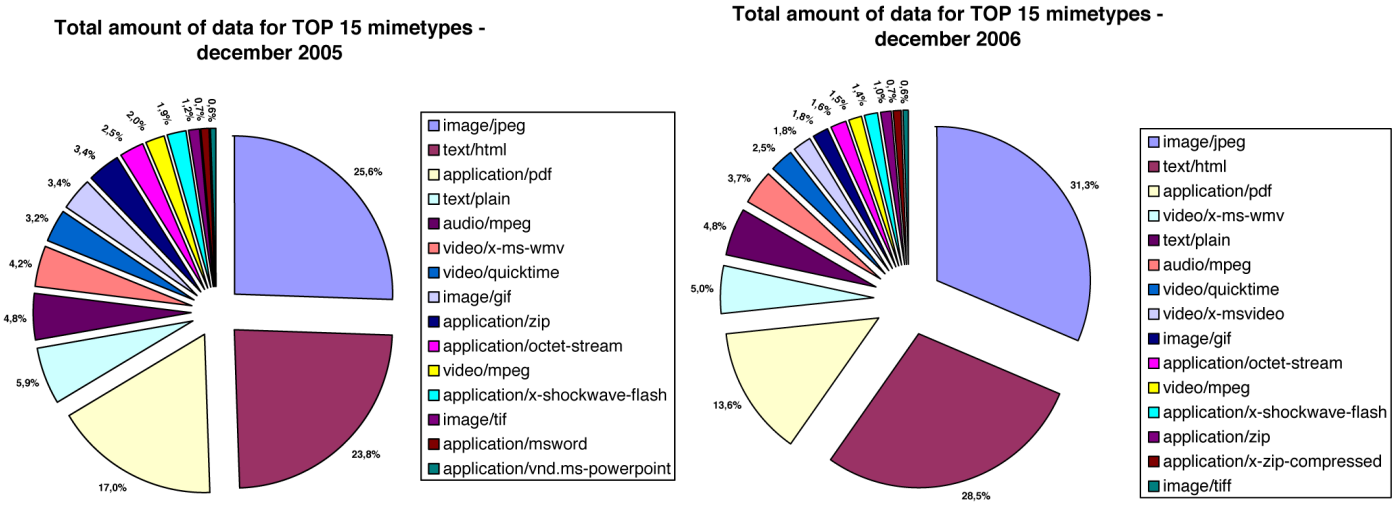
Dominios

La primera lista de dominios daneses de junio de 2005 contenía aproximadamente 607.000 nombres de dominio, de los cuales unos 480.000 estaban activos dentro del dominio <.dk>, mientras que la lista más reciente, de marzo de 2007, contenía en torno a 803.000 nombre de los cuales 640.000 estaban activos. La mayoría de los dominios (más del 85% de los dominios activos) siguen siendo de tamaño pequeño, es decir, menos de 10 MB. Las Webs más grandes pertenecen a dos corporaciones nacionales de radiotelevisión, la pública DR y la semiprivada TV2, que suman entre ambas más de 200 GB -y si incluimos material derivado, varios TB. A 1 de marzo de 2007, el archivo contiene 28,350 GB recopilados por medio de fotos fijas; 3,593 GB procedentes de capturas selectivas y 3,316 GB procedentes de capturas de acontecimientos.

Tipos de documentos

La última captura de fotos fijas revela que las imágenes (archivos jpeg) siguen siendo el tipo de archivo más común, representando en la actualidad más del 31% del total de datos recopilados (frente al 25% en 2005). En segundo lugar –lo que no es tan sorprendente– se sitúan los archivos de texto/html (28% en 2007, 23% en 2005). El tercer tipo son los PDF, cuya proporción en el total se ha reducido (13,6% en la actualidad, frente a 17% en 2005). Los archivos video/x-ms-wmv sustituyen a los archivos texto/sencillos en el cuarto lugar y el video en general está ahora representado con cuatro tipos de archivos entre los 15 tipos de archivos más comunes, frente a tres tipos en 2005. En términos

menos técnicos, esto significa que la población danesa utiliza de manera generalizada Internet para almacenar fotografías privadas y videos domésticos.



***Leyenda de los gráficos anteriores*:**

Volumen total de datos de los PRIMEROS 15 Mime Types **Diciembre 2005**	**Volumen total de datos de los PRIMEROS 15 Mime Types** **Diciembre 2006**
imagen/jpeg	imagen/jpeg
texto/html	texto/html
aplicación/pdf	aplicación/pdf
texto/sencillo	video/x-ms-wmv
audio/mpeg	texto/sencillo
video/x-ms-wmv	audio/mpeg
video/quicktime	video/quicktime
imagen/gif	video/x-msvideo
aplicación/zip	imagen/gif
aplicación/octet-stream	aplicación/octet-stream
video/mpeg	video/mpeg
aplicación/x-shockwave-flash	aplicación/x-shockwave-flash
imagen/tif	aplicación/zip
aplicación/msword	aplicación/x-zip-compressed
aplicación/vnd.ms-powerpoint	imagen/tiff

Problemas encontrados

Preguntas/quejas

Hemos recibido en torno a 100 preguntas y quejas. Teniendo en cuenta el número de Webs activas que hemos visitado (más de 650.000), se trata de una cifra muy baja. Cabe señalar que siempre dejamos nuestra “tarjeta” identificando al responsable de la captura y con una dirección de correo electrónico que se comprueba todos los días. Al principio, la mayoría de las preguntas se referían al hecho de que no respetábamos la convención robo.txt, lo que es cierto, ya que queremos capturar

todas las páginas públicas. Prácticamente todas las personas que han remitido quejas se han mostrado satisfechas cuando han sido informadas de que el propósito era el depósito legal. En este momento, sólo hay tres propietarios de Webs que siguen sin estar satisfechos con nuestra captura.

Crawler traps (Trampas para la araña)

Cuando nuestro personal rastrea la Web a través de enlaces, siempre termina cayendo en "*crawler traps*" (trampas para la araña), como por ejemplo calendarios, con un número infinito de enlaces. Muchas de estas trampas se descubren y se señalan, o bien en el dominio en el que han sido encontradas o en un nivel global, evitándolas en otras Webs que encierren la misma trampa.

Contraseñas

Los sitios que requieren una contraseña siguen siendo un problema, al menos por lo que respecta a la recopilación automática de información de acceso (login) (por ejemplo, la notificación automática por correo electrónico cuando se descubre el login generaría correo spam en muchas Webs no públicas que no están cubiertas por la ley y que no deberían archivarse). El responsable de la captura puede hacer el login técnicamente cuando está efectuando la captura pero la captura de información de login sigue siendo un proceso manual.

Organización administrativa

La administración de la ley de depósito legal danesa es responsabilidad conjunta de dos bibliotecas de depósito legal, la Royal Library y la State and University Library. Para asegurar una administración eficaz de la sección que se ocupa de "Materiales publicados en la red de comunicaciones electrónicas", se constituyó una institución virtual denominada "netarkivet.dk" (netarchive.dk). Esta institución está gestionada por un Comité de Dirección compuesto por 6 miembros (3 de cada biblioteca) con experiencia en Web, TI, depósito legal y desarrollo de colecciones. El Comité se reúne 2-3 veces al año, en función de las necesidades, para discutir y decidir sobre cuestiones económicas, técnicas, políticas y legales. Netarkivet.dk cuenta con una persona encargada de la gestión diaria que depende del Comité de Dirección y que se encarga de la supervisión del trabajo diario de netarkivet.

Para prestar asistencia a netarkivet en materia de formulación de políticas para el desarrollo de colecciones, el Ministerio de Cultura ha designado un Consejo Asesor Editorial compuesto por cinco miembros del ámbito universitario, TI y del sector editorial, siendo todos ellos usuarios actuales y futuros y socios colaboradores en el establecimiento de un archivo de red nacional. La primera convocatoria de este Consejo fue en marzo de 2006 y desde entonces se ha reunido dos veces con representantes de netarkivet.dk. El Consejo ha sido de gran ayuda a la hora de discutir y formular políticas sobre captura, definir "acontecimientos" y ayudar a identificar sitios que deben ser objeto de una captura selectiva.

Políticas

Junto con el Consejo Asesor Editorial, hemos formulado políticas sobre captura de acontecimientos (ver más arriba) y sobre dominios al margen de <.dk> que deben ser recopilados.

Dominios al margen de <.dk>

Como la ley también cubre material en Internet "publicado en otros dominios de Internet, etc. y que se dirige al público danés", también recopilamos Webs fuera del dominios <.dk> que se considera están "dirigidas al público danés", fundamentalmente Webs en danés, pero también Webs de compañías danesas, artistas e instituciones con sede social en Dinamarca. Hemos encontrado en torno a 38.000 Webs, cubiertas por la ley, la mayoría de ellas en <.com>. Los métodos para localizarlas han sido parcialmente automáticos y parcialmente manuales. El método automático consistía en cotejar una lista de enlaces incluidos en Webs <.dk> que dirigían fuera del dominio (y por tanto no capturados) con un GEO-IP que supuestamente revela Webs con propietarios daneses. El método manual consistía en buscar en Google páginas Web que incluían nombres de lugares daneses pero que no estaban en el dominio <.dk>, lo que permitió llegar a muchas Webs de instituciones y asociaciones locales.

Acceso

Lamentablemente, de momento, el acceso es extremadamente limitado. Esto se debe no a la legislación sobre copyright, como pudiera esperarse, sino a la Ley de Tratamiento de Datos de Carácter Personal. La Agencia Danesa de Protección de Datos ha decidido que la información recopilada puede contener datos personales sensibles y, por consiguiente, el acceso a netarchive debe ser muy restrictivo. En la práctica, esto significa que sólo los investigadores involucrados en un proyecto de investigación de nivel postdoctorado están autorizados a acceder. Evidentemente, esto es contrario a la filosofía subyacente a la Ley de depósito legal, a saber, que, una vez publicados, los materiales están disponibles públicamente en la Biblioteca Nacional. Por tanto, el Ministerio de Cultura va a negociar con la Agencia de Protección de Datos para que se permita un acceso más amplio.

Personal y Equipos

Netarchive cuenta con el equivalente a 4,5 empleados a tiempo completo/año, distribuidos entre unos 20 empleados, ingenieros de TI, informáticos, bibliotecarios y ayudantes de biblioteca.

El sistema técnico completo consta actualmente de 9 servidores Linux (administración, captura y acceso) y 25 PC con Windows XP (para indización y preservación de bit). En nuestra Web: http://netarchive.dk pueden encontrar más artículos técnicos con la descripción del sistema, tanto a nivel general como artículos con detalles muy técnicos y específicos.

¿Más preguntas?

Consulte www.netarchive.dk o póngase en contacto con Grethe Jacobsen, gja@kb.dk o con Eva Fønss-Jørgensen, efj@statsbiblioteket.dk (para cuestiones legales) o con Bjarne Andersen bja@netarkivet.dk (para cuestiones técnicas).