Congreso Mundial de Bibliotecas e Información:
71 Congreso General y Consejo de la IFLA

"Libraries - A voyage of discovery"

14 – 18 de agosto de 2005, Oslo, Noruega

*Conference Programme:* http://www.ifla.org/IV/ifla71/Programme.htm

*July 18, 2005*

**Código Número :** 074-S
**Reunión :** 133 SI - Bibliografía

# El buscador de Web (*Web crawling)*: La experiencia de la Bibliothèque nationale de France

**Christian Lupovici**
Jefe de la Agencia Bibliográfica Nacional
Bibliothèque nationale de France
Christian.lupovici@bnf.fr

**Traducción al español: Francisca Movilla López**

***Resumen:***
*La Bibliothèque nationale de France en relación con su misión sobre el Depósito legal está experimentando actualmente métodos de recolección de web y está organizando la preservación a largo plazo de los documentos digitales. El trabajo realizado para lograr este objetivo incluye reflexiones básicas sobre los fundamentos del Depósito legal y el tratamiento bibliográfico de los recursos de Internet. Al trabajar de un modo real, la gran cantidad de recursos digitales es un factor importante que ayuda a la Biblioteca Nacional a tomar cualquier decisión.*

## 1. El Depósito legal en la BnF

### *1.1 Antecedentes históricos*

La legislación francesa sobre Depósito legal se remonta al siglo XVI cuando Francisco I dictó la Ordenanza de Montpellier en 1537. A partir de esa fecha, en todas las regulaciones posteriores se ha procurado tener en cuenta la aparición de nuevas tecnologías para integrarlas en la normativa legal, haciendo cada vez más amplio el ámbito del depósito legal.
*1537: libros;*
*1667: grabados;*
*1689: estampas;*
*1925: toda producción de arte gráfico;*
*1941: carteles, partituras y fotografía;*
*1963: registros sonoros de todo tipo;*
*1975: registros de películas fijas o animadas, cualquiera que sea su proceso técnico;*
*1977: obras de cine;*

*1992: Todo tipo de documentos media, incluyendo bases de datos y sistemas expertos cuando sean para distribución al público.*

La inclusión de los recursos web en la normativa francesa se realizará en 2005, ya que el tema está siendo debatido en el Parlamento.

### *1.2 Misión y alcance del Depósito legal*

Está claro que el objetivo del Depósito legal es preservar y permitir el acceso a largo plazo a todo el patrimonio cultural de documentos producidos y distribuidos en gran número de copias y dar acceso a ese patrimonio. El Depósito legal incluye todo tipo de recursos y no se limita únicamente a los de lengua francesa, sino a todos los documentos publicados en el territorio nacional.
Es más una cuestión del tipo de contenido y/o canales de distribución (libros, estampas, mapas, registros sonoros, audiovisuales) que una cuestión de soporte: papel, disco o en línea. En Internet, la mayor parte de la literatura gris institucional estará incluida en la regulación del Depósito legal siempre que los documentos en la web sean accesibles al público, por lo tanto publicados.
El Depósito legal supone además que la recolección de documentos no es una cuestión de adquisición, en el sentido de una política de desarrollo de la colección, sino una cuestión de recogida de material de todo tipo que dé una idea de los contenidos distribuidos al público francés en el momento actual, con el fin de preservarlos para estudios históricos futuros. Esto supone que están incluidas todas las versiones diferentes de un documento, así como las obras prohibidas, como ha ocurrido siempre desde el siglo XVI.
¿Cómo trasladar esto a los recursos Web?

## 2. El buscador de Web (*Web crawling)*

### *2.1 La recolección en Internet con fines de Depósito legal*

La transposición de esta filosofía del Depósito legal a los recursos Web implica:
a) localizar todos los nombres de dominio que no contienen claramente recursos pertenecientes a la cultura francesa;
b) centrar la atención en los nombres de dominio específicos del ámbito nacional francés (*.fr, .pf, .wf, .pm, .re, .tf, .ad, .yt*);
c) recopilar todo lo demás.

En una primera aproximación, no es posible saber lo que es relevante para la cultura francesa dentro de "lo demás" y hay que trabajar con el fin de encontrar el mecanismo para hacer una selección aproximada, basada en el análisis semántico del contenido de las sedes Web. Sin embargo, como el espacio para el almacenamiento ya no es un problema crucial, la Bibliothèque nationale de France, está guardando y preservando una serie de sedes en mayor número de lo que sería estrictamente necesario con fines de Depósito legal, en espera de mejores soluciones. En todo caso, la parte de la web que se guardará y preservará siempre será grande.

### *2.2 De los experimentos al plano real*

#### 2.2.1 Primeros experimentos

La Bibliothèque nationale de France lleva recolectando sedes web francesas mediante el buscador HTTrack desde 2001[1]. Los primeros experimentos se centraron en varias campañas para las elecciones en Francia. Estos experimentos pretendían comprobar diferentes puntos:
a) cómo construir una tipología de sedes web. Y de acuerdo con ello, cuál es la periodicidad correcta para aplicar el buscador;
b) comprobar los problemas técnicos (formatos, exhaustividad en la búsqueda) y puesta a punto del buscador;
c) evaluar los mecanismos manuales y automáticos en la selección de las sedes web.

Esta fase era importante para valorar la selección automática de las sedes web para procesarlas manualmente y cómo utilizar instrumentos que ayudaran a la toma de decisiones.

[1] http://bibnum.bnf.fr/conservation/aristote2004/roche.pdf

### 2.2.2 El buscador de Web (*Web crawling*) en el plano real

¿Qué es lo que se recoge en el Depósito legal francés en cuanto a recursos y volumen?
Se hizo un extracto de la panorámica realizada sobre el conjunto de la web en la primavera de 2004 por el buscador de web Alexa, según las especificaciones de los dominios en los que había que buscar (*.fr, .com, .net, .edu, .com, .biz, .info, .int, .pf, .ad, .coop, .name, .aero, .tf, .re, .museum, .pro, .pm, .wf, .yt*) y los que se excluía. Del resultado de este rastreo se extrajeron más de 2 billones de URLs y 65 millones de "hosts", con un peso de 27 terabytes.
Además, se implantó un buscador centrado en el domino *.fr* a finales de 2004 y en enero de 2005 utilizando el buscador Heritrix. El resultado de este rastreo posterior aportó 500 000 sedes públicas y 4 000 sedes personales. Esto supone más de 118 millones de URLs y medio millón de "hosts", con un peso de 3 terabytes.
Para este volumen de datos es impensable cualquier proceso de catalogación o indización manual más que algo complementario. Tal cantidad de datos necesitaría un mínimo de 500 personas/año para clasificar y catalogar solamente las sedes públicas. Esto es dos veces más que la plantilla de catalogación de la Bibliografía nacional en la BnF.
Por ello es necesario combinar diferentes técnicas.

## *2.3 El proceso de Depósito legal de la Web*

El método fundamental es un buscador automático. Toda la web de superficie y parte de la web profunda puede ser recogida de esta forma, dependiendo de los problemas técnicos a los que se tenga que enfrentar el buscador y de cómo se haya definido. Pero en líneas generales, una búsqueda automática recogerá principalmente la web de superficie, creando un mapa de la web.
Un segundo método complementario es centrar la búsqueda automática en un conjunto de sedes específicas, por ejemplo, en el dominio *.fr*, y pedir al buscador que realice una recolección más a fondo.
Estos dos últimos métodos están orientados de arriba a bajo.
Con el fin de conseguir una panorámica completa de la web, debemos combinar la recogida automática periódica, con una técnica de orientación de abajo hacia arriba que consiste en complementar las colecciones con los recursos de la web profunda depositados por productores de sedes y reproducir los enlaces con la web de superficie recogida anteriormente. Estos depósitos tendrán que ser procesados en parte de forma manual y en parte automáticamente. El depósito puede presentar ventajas en cuanto a la seguridad de preservación, pero es un proceso lento, empezando por el análisis de la sede, que precisa 5 personas/día como media.
En la BnF el conjunto de datos se almacena en un repositorio de datos central para uso interno y utilización del personal. El archivo web es parte del programa global de la BnF de almacenamiento y preservación digital, SPAR (*Système de Préservation des ARchives*) puesto en marcha a finales de 2004.

## *2.4 Cuestión y método de selección*

En principio, a la Bibliothèque nationale de France le gustaría archivar la mayor parte de las sedes web relevantes. Por eso, es interesante encontrar un método que pueda seleccionar o proponer una selección entre todas las sedes recolectadas. Se llevo acabo un proyecto de investigación francés llamado "Watson" para aplicar tratamientos lingüístico en el archivo web de la BnF por usuarios y personal de plantilla. El proyecto duró de 2003 al 2004[2]. Una de las formas aplicadas consistía en la caracterización de la sede realizando el análisis lingüístico de los metadatos y de las frases incluidas en el texto completo. Este método proporcionó resúmenes sobre el contenido de la sede. Aún no es lo suficientemente bueno para usarse sin una comprobación humana, pero está bastante bien como instrumento de ayuda para la selección de sede.
La BnF está organizando un equipo de observadores de la web formado por profesionales que trabajan en distintos departamentos de la Biblioteca, principalmente encargados del depósito legal. El objetivo del equipo es aprender a gestionar este nuevo contexto de recursos digitales y realizar la conexión necesaria con las colecciones tradicionales anteriores sobre soporte. Se han creado las herramientas de acceso y los métodos de observación para evaluar el resultado de la

[2] José Coch, Julien Masanès**. -** Language engineering techniques for web archiving. In : 4th International Web Archiving Workshop (IWAW04), 16 septembre 2004, Bath, UK - http://www.iwaw.net/04/index.html (visited 30/05/2005).

búsqueda y la calidad de la recolección. El equipo cuenta con un instrumento que les permite hacer verificaciones tanto en el archivo de la BnF como en la web, con el fin de señalar aquellas sedes web en las que hay que rastrear en profundidad mediante un buscador con un objetivo claro. Este otoño, el buscador tendrá en cuenta el trabajo realizado por el equipo de observadores y propondrá las URLs "preseleccionadas" incluidas en sus parámetros.
En todo caso, nunca se podrá seleccionar el conjunto de sedes web que correspondan exactamente al Depósito legal francés y es evidente que el archivo web de la BnF será siempre de gran tamaño. Si todos los países están haciendo lo mismo habrá un solapamiento, pero es una solución mejor dado que el almacenamiento ya no es un problema y la existencia de duplicados en distintos países dará una mayor seguridad para la preservación a largo plazo. Por otro lado, la técnica panorámica (de visión global) no asegura la exhaustividad de las colecciones sino un ejemplo representativo de la cultura actual. Este es el fundamento del Depósito legal.
Sin embargo, el consorcio internacional IIPC (International Internet Preservation Consortium) reconoce la necesidad de un "buscador inteligente" capaz de centrarse en conjuntos de sedes de acuerdo con unos parámetros fijados y capaz de profundizar cuando el nivel de relevancia es alto. La British Library y la Bibliothèque nationale de France han hecho un llamamiento, esta primavera, para la presentación de propuestas de un buscador inteligente.

### *2.5 Nuevo contexto, nuevo documento*

Al gestionar la web debemos abandonar la forma de proceso del documento tradicional y avanzar hacia una nueva forma de gestión de los recursos.
El archivo web tiene nuevas dimensiones:
Debe mantener la posibilidad de navegar que se regenera internamente. Ningún documento es un recurso independiente. El propio documento se está transformando. En el modelo de la BnF la unidad documental es la propia sede web (en un modelo de colección).

### *2.6 Técnicas lingüísticas para el análisis de recursos*

Frente a la gran cantidad de datos acumulados en un sistema de almacenamiento, tanto el profesional como el lector necesitan instrumentos para buscar la información que necesitan. Puede que necesiten reunir recursos en subgrupos para buscar después una información concreta, pero también puede que necesiten procesar una gran cantidad de información para aplicar tratamientos estadísticos o lingüísticos.
El proyecto Watson tenía como objetivo seleccionar grupos de recursos correspondientes a un dominio señalado y aplicar sobre este conjunto identificado algún análisis lingüístico.
Puede dar una idea del contenido informando al usuario sobre la estructura, palabras claves, nombres y localizaciones, etc. embebidos en los documentos o aportando un resumen en el que poder navegar.
El proyecto aportaba instrumentos específicos al usuario para la extracción temática, navegación contextual y análisis semántico del contenido de las sedes. Los experimentos se llevaron a cabo en el archivo de elecciones francesas.

## 3. La nueva catalogación del futuro

### *3.1 Catalogación de metadatos*

Tradicionalmente los metadatos se han utilizado para describir e identificar objetos físicos con miras a la recuperación. Esta actividad catalográfica supone una tarea de trabajo intensivo y totalmente manual, pero se adapta a los objetos físicos y no se puede evitar en este contexto.
En un medio digital la situación es diferente:
Los objetos digitales están (o deberían estar) preparados para procesar la información. Contienen la propia documentación que puede ser metadatos técnicos y legales que permiten su tratamiento en un sistema.
Cuando el texto está asociado, el propio contenido puede ser indizado y clasificado sin ninguna intervención humana (en este momento). Esto conlleva la información de estructura y del contenido semántico.
En el contexto de la web, si los metadatos descriptivos son escasos se compensan por efecto de la red. Los objetos digitales no son documentos aislados en el archivo web. Están vinculados a otros recursos mediante enlaces explícitos, contenido y descripción de contenido, que incluyen fuentes secundarias ya sean tradicionales o totalmente nuevas.

Esto no significa que ya no necesitemos los metadatos descriptivos o palabras clave. Pero la biblioteca no es el mejor lugar para elaborar esta información, es mejor trasladar esta tarea a la fase de creación, ya que la necesidad de este tipo de datos es menos importante que en un contexto tradicional.

### *3.2 Otros tipos de metadatos*

Están surgiendo nuevos problemas, como son la gestión de derechos, la garantía de autenticidad, no sólo por motivos legales, sino también para asegurar la reproducción ilimitada de los recursos digitales y su inmediata circulación. Los formatos, que a veces se combinan en un único documento, son uno de los aspectos cruciales, ya que la información del servidor o del documento no es suficientemente fiable. Los documentos recogidos necesitan una valoración del formato.
Dentro de una Biblioteca Nacional es además muy importante ampliar los metadatos a la información de preservación adaptada al archivo web.

### *3.3 Futuro de la catalogación*

La situación de la catalogación en el contexto web puede abreviarse mediante la realización de especificaciones para la creación del documento digital y el análisis del recurso en la fase de incorporación (modelo OAIS), incluyendo las instrucciones del buscador y la interrelación con los productores de web.

## 4. ¿Cómo puede ser una Bibliografía Nacional de la web?

La ley de Depósito legal francesa (incluso la nueva) señala la existencia de una Bibliografía Nacional elaborada a partir de los documentos depositados.
Pero en este contexto ¿cuál es la apariencia de una Bibliografía Nacional?
Dado que la finalidad de la Bibliografía Nacional es difundir las nuevas publicaciones disponibles, podemos plantearnos el publicar en la sede web de la BnF, la lista de las nuevas sedes aparecidas (no las modificaciones), aplicando un tratamiento de clasificación y "catalogación" automáticas a la fase de archivo (modelo OAIS). El resultado será una lista de URLs con índices de nombres, de materias (hasta donde sea posible hacer esto con un análisis lingüístico, parecido a lo que se ha hecho en el proyecto Watson).