Consejo de Cooperación Bibliotecaria

Grupo de Trabajo de Patrimonio digital

Consejo de Cooperación Bibliotecaria

Informe (enero, 2013)

Coordinadores:
Eugènia Serra. Biblioteca de Catalunya
Mar Pérez Morillo, Biblioteca Nacional de España

Consejo de Cooperación Bibliotecaria

Informe del Grupo de Trabajo
Patrimonio digital

1. Miembros del grupo

Coordinadores:

Comunidad Autónoma de Cataluña
Eugenia Serra Aranda
Biblioteca de Catalunya

Mar Pérez Morillo
Biblioteca Nacional de España

Participantes:

Comunidad Autónoma de Andalucía
Jesús Jiménez Pelayo
Biblioteca de Andalucía

Comunidad Autónoma de Aragón
Carmen Lozano Floristán
Biblioteca de Aragón

Comunidad Autónoma de Cantabria
José Mª Gutiérrez Rodríguez
Servicio de Archivos y Bibliotecas

Comunidad Autónoma de Castilla y León
Alejandro Carrión Gútiez
Biblioteca de Castilla y León/BPE en Valladolid

Comunidad Autónoma de Extremadura
Leonor Gómez Barroso
Biblioteca de Extremadura

Comunidad Autónoma de Galicia
Noelia Bascuas Ferreiro
Biblioteca de Galicia

Comunidad Autónoma del País Vasco
Francisca Pulgar Vernalte
Servicio de Bibliotecas. Dirección de Patrimonio Cultural

Comunidad de Madrid
Luisa Inmaculada Fernández Miedes
Coordinación Técnica de la Subdirección General

Comunidad Foral de Navarra
Juan Francisco Elizari Huarte
Biblioteca General de Navarra

Comunidad Valenciana
Mª Rosario Tamarit Rius
Biblioteca Valenciana

Principado de Asturias
Milagros García Rodríguez
Biblioteca de Asturias "Ramón Pérez de Ayala"

Región de Murcia
Fina Sánchez Fernández
Biblioteca Regional de Murcia/BPE en Murcia

Ciudad Autónoma de Ceuta
Ana Leria Ayora
Biblioteca del Museo de Ceuta

Ministerio de Educación, Cultura y Deporte. Secretaría de Estado de Cultura
Isabel García-Monge Carretero
Servicio de Fondos Especiales. Subdirección General de Coordinación Bibliotecaria

2. Objetivos 2012

En la reunión de Santander se definieron como objetivos del Grupo para el 2012:

1. Definir el tipo de recursos electrónicos web susceptibles de ser capturados para su preservación
2. Coordinarse con el GT de Depósito legal en la regulación relativa a Depósito Legal.
3. Abrir una vía de comunicación con el sector de archivos para saber cómo están abordando el tratamiento de la documentación electrónica que generan empresas, instituciones y administraciones.

3. Actividad del Grupo

Objetivo 1: ***Definir el tipo de recursos electrónicos web susceptibles de ser capturados para su preservación.*** El informe del año anterior incluia una primera lista de recursos según tipología, editor/productor y temática. A lo largo del 2012 no se ha trabajado dicha lista más al detalle, dado que dependía en gran medida de lo

que estableciera el nuevo Real Decreto sobre DL, que se estaba redactando en el Grupo de DL.

Objetivo 2: Coordinarse con el GT de Depósito legal en la regulación relativa a Depósito Legal. Se ha informado al Grupo de los trabajos del GT de Depósito Legal. Además, los representantes de Aragón, Asturias, Cantabria, Castilla-León, Catalunya y Navarra, al formar parte de ambos grupos, hemos servido -en cierta medida- de puente al conocer la opinión de unos y otros.

Objetivo 3: Abrir una vía de comunicación con el sector de archivos para saber cómo están abordando el tratamiento de la documentación electrónica que generan empresas, instituciones y administraciones. No se ha abordado esta línea de trabajo a la espera del redactado del Real Decreto. En todo caso, se trataba de un objetivo de una relevancia menor puesto que no dejan de ser casuísticas diferentes, en cuanto que la información del sector de archivos no es habitualmente pública en la Red, y nuestro marco de actuación es la web y los recursos públicos, ya sean gratuitos o de pago.

También en el informe anual pasado acordamos que anualmente recogeríamos la **información actualizada de las inciativas de recogida de la web en curso**, dicha información se encuentra en forma resumida en el punto siguiente, además de anexar los informes remitidos por Euskadi, BNE y Catalunya.

4. Información actualizada de las iniciativas de recogida de la web en curso

En los anexos 1, 2 y 3 se adjunta la información completa proporcionada por el País Vasco, Catalunya y la Biblioteca Nacional de España.

Cada informe describe el software y hardware utilizado, los criterios de captura, cuantificación del patrimonio recolectado y en el caso de Catalunya y Euskadi la infraestructura de captura y almacenamiento.

El informe de la Biblioteca Nacional de España aporta datos de los criterios de captura, periodicidad, software y hardware para el almacenamiento. No incluye la infraestructura de captura específica puesto que utiliza la plataforma de Internet Archive que es quién se encarga de realizar la recogida.

A continuación se incluye un resumen con los datos principales extraídos de los tres informes.

RESUMEN DE DATOS DE LOS INFORMES

<table>
<tr><th></th><th>PADICAT (CATALUNYA)</th><th>ONDARENET (EUSKADI)</th><th>BIBLIOTECA NACIONAL DE ESPAÑA</th></tr>
<tr><td rowspan="4">SOFTWARE</td><td>Heritrix para la captura</td><td>Heritrix para la captura</td><td>Heritrix para la captura</td></tr>
<tr><td>Nutchwax, Hadoop y Wayback para la indización</td><td>Nutchwax para la indización</td><td>Lucene para la indización</td></tr>
<tr><td>Wera y Wayback para la búsqueda</td><td>Wayback para la búsqueda</td><td>Wayback para la búsqueda</td></tr>
<tr><td>Web Curator para metadatos y gestión</td><td>Web Curator para metadatos y gestión</td><td></td></tr>
<tr><td rowspan="3">HARDWARE</td><td>4 nodos HP Proliant DL360 G4p/X3,0 con 8 GB RAM para la captura</td><td rowspan="2">Entorno de Pruebas:
1 servidor con características similares a las de un HP ProLiant DL380 G5 con: Procesador Intel® Xeon® a 3 GHz, Dual Core, 2 Procesadores, 12 MB (2 x 6 MB) de caché de nivel 2; 4 GB (4 x 1 GB) de memoria de serie; controlador HP Smart Array P400/512 MB con caché de escritura respaldada por batería (RAID 0/1/1+0/5/6); ampliación Memoria: 4GB FBD PC25300 (2x2GB); disco SAS SFF de 146GB 2.5 Conectable en caliente 10.000 rpm</td><td rowspan="3">PetaBox, una unidad de almacenamiento y procesamiento de datos diseñada por el Archivo de Internet y Saikley CR</td></tr>
<tr><td>1 nodo HP Proliant DL360 G5/X5110 con 8 GB RAM para indización</td></tr>
<tr><td>1 nodo dHP Proliant DL360 G5/X5110 con 16 GB RAM). 1 clúster Linux de balanceo. Para búsqueda y visualización</td><td>Entorno de Producción:
3 servidores con características similares a las de un HP ProLiant DL380 G5.</td></tr>
</table>

	PADICAT (CATALUNYA)	ONDARENET (EUSKADI)	BIBLIOTECA NACIONAL DE ESPAÑA
	1 cabina NetApp FAS3170 de capacidad de 19 TB para almacenamiento	Sistema de archivos (filesystem) dividido en 2: uno al 87% y el otro al 81%. El resultado sería la media de los 2: 84%. Cuando el primer filesystem llegue al 90% las descargas programadas se empezarán a descargarse en el segundo.	
CONTENIDO	249.609 URLs	16.943 URLs	
	349.000.000 archivos	25.195.236 archivos	1.780 millones archivos
	13 TB espacio	824,04 GB espacio	85 TB
POLÍTICA DE COLECCIÓN	30.000 recursos del dominio .cat. 2 capturas/año	Diarios digitales: una o dos descargas semanales de un perfil bajo (10 MB)	1 recolección/año masiva de recursos dominio .es y subdominios asociados (.com; .edu; .gob; .net; .biz; .info; .org)
	578 webs de 450 entidades con convenio. 2 capturas/año	Revistas digitales: captura según su periodicidad, de un perfil bajo (10 MB)	
	800 webs recomendadas. 2 capturas/año	Webs: 1 captura/año (portada y 1 o 2 niveles de profundidad)	
	1 recursos de monográficos. 1 captura/año	Blogs, captura de periodicidad variable y perfil bajo (cada 6 meses o anual)	1 recolección selectiva de Humanidades
	30 publicaciones en serie. Captura diaria		

5. Conclusiones

Teniendo en cuenta que:

a) una parte de los objetivos del GT de Patrimonio Digital son comunes o dependientes de los del GT de Depósito Legal,
b) una parte de los miembros son comunes a ambos grupos,

y una vez consultados los miembros del Grupo, así como a la coordinadora del GT de Depósito Legal, se propone la fusión del Grupo de DL y del Grupo de Patrimonio Digital (posibilidad que ya se avanzó en la reunión de Santander del CCB). Esta fusión permitiría concentrar esfuerzos en vez de disgregarlos.

En cuanto a la forma de proceder, se podría actuar de la siguiente manera:

a) modificar el nombre del GT de Depósito Legal añadiéndole "y de Patrimonio Digital",
b) cerrar el GT de Patrimonio Digital,
c) ofrecer a os miembros del GT disuelto que si lo desean, se incorporen al GT de Depósito Legal y Patrimonio Digital.

ANNEXO 1

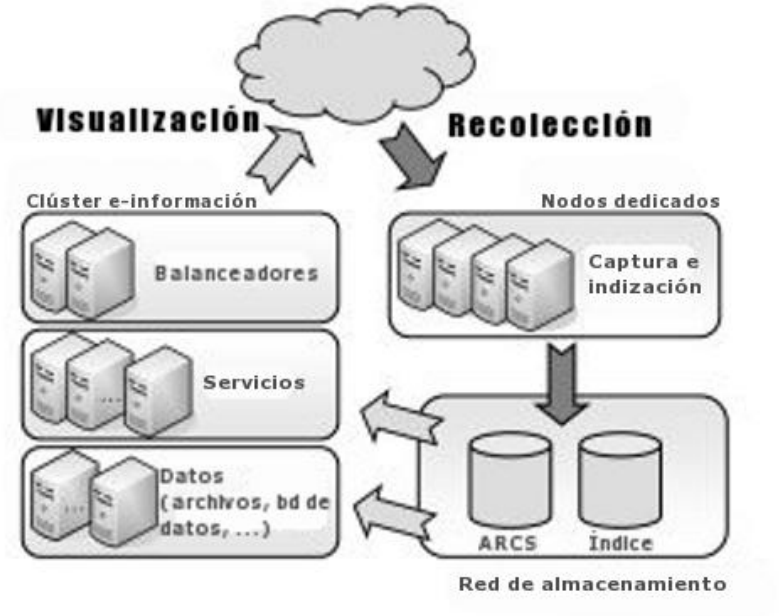
PADICAT (Patrimonio Digital de Cataluña)
Estructura técnica y capacidades (enero 2013)

Hardware

La estructura descrita a continuación se encuentra ubicada en el Centre de Serveis Científics i Acadèmics de Catalunya (CESCA).

- 4 nodos para la captura (HP Proliant DL360 G4p/X3,0 con 8 GB RAM)
- 1 nodo de indización (HP Proliant DL360 G5/X5110 con 8 GB RAM)
- 1 nodo de servicios para la capa web, con las interfaces de búsqueda y visualización wayback y wera (HP Proliant DL360 G5/X5110 con 16 GB RAM)
- 1 clúster Linux de alta disponibilidad con características de balanceo de carga de peticiones y de tolerancia de errores en caso de desastre técnico de los nodos que integran la plataforma
- 1 cabina NetApp FAS3170 que presenta en un espacio de disco vía NFS los nodos anteriores, de capacidad de 19TB.
- 1 robot donde se conservan las copias de seguridad de los datos en cinta.

Además, la Biblioteca de Catalunya (BC) tiene un repositorio seguro de preservación llamado COFRE (COnservamos para el Futuro Recursos Electrónicos), ubicado en la propia BC y desarrollado por la propia BC, en el que se cargarán una o dos copias anuales del contenido de PADICAT, junto con los demás recursos digitales que preserva la BC.

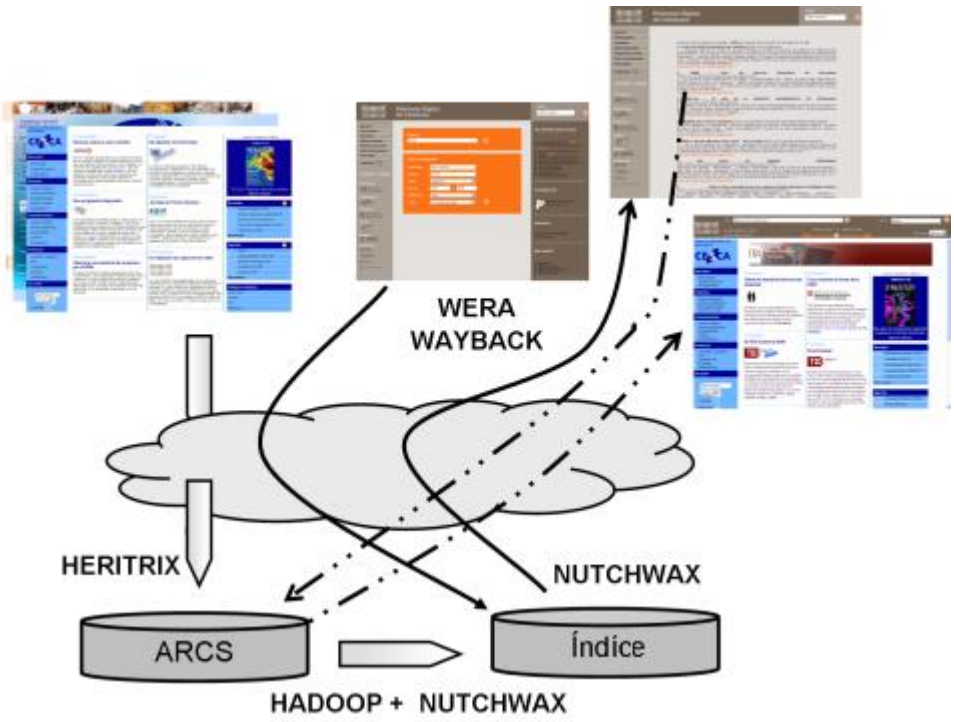


Software

Se trata de software de código abierto y gratuito, que se ha desarrollado dentro de la comunidad internacional, principalmente bajo el paraguas del IIPC.

- Heritrix compila las páginas web tal como las ve el usuario que navega por Internet, y als almacena en ficheros comprimidos ARC.
- A continuación, Heritrix se complementa con NutchWax, o con la combinación de Hadoop y Wayback, para realizar los procesos de indización.

- Wera permite la búsqueda a texto completo en los índices generados por NutchWax.
- Wayback permite la búsqueda por URL en los índices generados por Hadoop y el mismo Wayback.
- Web Curator Tool se ha integrado como sistema para asignación de metadatos y otras tareas de gestión de colección.



Contenido

58.122 páginas webs diferentes identificadas por URL
249.609 capturas de versiones de las páginas web
349 millones de ficheros informáticos
13 TB de espacio en disco

Política de colección

- Compilación semestral de 30.000 recursos del dominio .cat
- Compilación semestral de 578 sitios web de 450 entidades que han firmado convenio con la BC (ayuntamientos, universidades, partidos políticos y sindicatos, clubs deportivos, sociedades culturales y museos, ONG, medios de comunicación, empresas, etc.)
- Compilación semestral de 800 páginas web recomendadas.
- Compilación anual de unos 1.000 recursos de monográficos.
- Compilación diaria de una parte substancial de 30 publicaciones seriadas en línea.

ANNEXO 2

EUSKAL ONDARE DIGITALA
ONDARENET
PATRIMONIO DIGITAL VASCO

HARDWARE

En cuanto al hardware en si:

- Pruebas, se dispone de 1 servidor con características similares a las de un HP ProLiant DL380 G5 con:
  - Procesador Intel® Xeon® a 3 GHz, Dual Core
  - 2 Procesadores,
  - 12 MB (2 x 6 MB) de caché de nivel 2
  - 4 GB (4 x 1 GB) de memoria de serie
  - Controlador HP Smart Array P400/512 MB con caché de escritura respaldada por batería (RAID 0/1/1+0/5/6)
  - Ampliación Memoria: 4GB FBD PC25300 (2x2GB)
  - Disco SAS SFF de 146GB 2.5 Conectable en caliente 10.000 rpm
- Entorno de Producción: 3 servidores con características similares a las de un HP ProLiant DL380 G5.

Tenemos un sistema de archivos (filesystem) que está dividido en 2: uno al 87% y el otro al 81%. El resultado sería la media de los 2: 84%. Cuando el primer filesystem llegue al 90% las descargas programadas se empezarán a descargarse en el segundo.

SOFTWARE

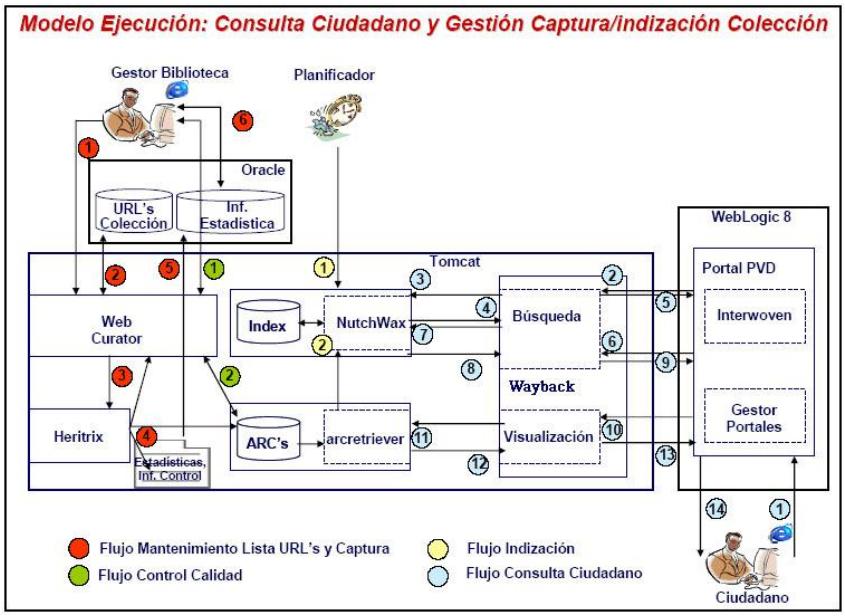
El entorno tecnológico elegido para lograr los objetivos, se basa en herramientas utilizadas en proyectos similares que Ondarenet y recomendados por la IIPC, y que se basan en un sistema de *código abierto*. El toolkit que se utiliza está formado por los siguientes programas:

- **Heritrix**: robot de captura de sitios y elementos web.
  - Realización de capturas/parámetros.
  - Recogida de datos estadísticos.
  - Almacena las capturas en ficheros ARC.
- **NutchWax**: indexación y búsqueda:
  - Indexa elementos: textos y documentos.
  - Permite indexar ficheros ARC (Wax).
  - Permite búsquedas no elimina duplicados.

- **Web Curator**: es el interface del gestor de la colección:
  - Gestión de los usuarios.
  - Define, configura y planifica las listas de capturas.
  - Define los metadatos.
  - Control de la calidad de las descargas.
- **Wayback**: interface de consulta del ciudadano:
  - Búsqueda de elementos.
  - Visualización de la historia-versiones.
  - Visualización de versiones.
- **Administrador de descargas**: Con la idea de gestionar las indexaciones y publicaciones se ha creado un pequeño portal para que el técnico documentalista haga un seguimiento del estado de las descargas y de las publicaciones de las mismas. De esta forma se pueden despublicar o publicar las descargas realizadas desde la herramienta Web Curator Tool. Esta función permite que muchas url's a las que sus autores no han dado autorización para su publicación en Ondarenet, se preserven y guarden dentro del servidor para cumplir con los objetivos de Ondarenet de guardar y preservar el patrimonio digital vasco.

Así mismo con el programa administrador nos permite acceder a otras funciones:

- Web Curator Tool.
- Ondarenet.



- Estadísticas.

PROGRAMACION DE DESCARGAS

En cuanto al número de descargas que se piensa realizar se tiene los siguientes criterios:

-Publicaciones periódicas: dependiendo del tipo de publicación:

- Diarios digitales: se ha modificado el criterio realizando una o dos descargas semanales, y de un perfil bajo (10 MB).
- Revistas digitales: según su periodicidad se aplican diferentes programaciones: semanal, dos veces a la semana, mensual,… , y de un perfil bajo (10 MB).

-Webs: una descarga anual. Si se observa que no ha habido variaciones se desecha, y si se observa que ha modificado el diseño o los contenidos se captura. Ahora disponemos de un perfil de 20 MB, lo que nos permite descargar la portada y 1 o 2 niveles de profundidad.

-Blogs: en la actualidad muchas asociaciones y grupos optan en vez de disponer de una web de crear blogs e incluso de las redes sociales (Facebook y twitter) debido a que es más fácil actualizar la información y no dependen de una empresa para realizarlo. La descarga de los blogs a veces suele ser complicada ya que muchas veces suelen descargar fuera de su dominio, y descarga contenidos que no tienen que ver con el fin de Ondarenet. Para ello se suelen poner un perfil bajo de descarga de 10 o 20 MB. Debido a la constante actualización de los blogs, las capturas también varían por que pueden ser: mensuales, cada 6 meses o anual.

Lo que hay que realizar también anualmente es una revisión de las urls, ya que muchas desaparecen o se modifican, por lo que también se modifican los targets y las programaciones de los mismos.

ALGUNOS DATOS

En la actualidad se tienen realizas las siguientes capturas de "*targets*" u objetos de descargas por grupos que hacen un total de 1767 *targets*:

- Arte (352).
- Ciencia y tecnología (72).
- Cultura (163).
- Economía y negocios (66).
- Educacion e investigación (60).
- Empresa (149).
- Euskera (121).
- Ocio y cultura (199).
- Política y gobierno (241).
- Salud (74).
- Sociedad (196).

- Sociedad de la información (78).

En cuanto a colecciones especiales que se denominan "Puntos de interés", tienen 503 *target* en total. En el año 2012 se han realizado los siguientes grupos:

- Elecciones al Parlamento Vasco (2012).
- Diáspora Vasca (2012).

En cuanto a la distribución general de la tipología de los recursos y formatos de las descargas que hemos realizado, podemos decir que a día de hoy 18 de enero de 2013 tenemos las siguientes estadísticas:

- Por volumen de descargas: 824,04 Gb indexados.
- Por número de documentos: 25.195.236 documentos.
- Capturas totales: 16.943 capturas.
- 4.567 capturas. capturas de versiones de las páginas web
- 2.087 urls seleccionadas.

ANNEXO 3

EL ARCHIVO WEB DE LA BNE.
Datos enero 2013

En 2009 la BNE inició el sistema de preservación de los contenidos digitales españoles albergados bajo el dominio nacional .es. Por la envergadura de la recolección a acometer -tarea para la que la BNE, en caso de haber querido llevarla a cabo en solitario, carecía de recursos humanos y técnicos suficientes- se contrataron los servicios profesionales de la Fundación Internet Archive, pionera en el archivo de contenidos publicados en Internet, misión que viene realizando desde 1996.

El archivo web del IA es el más grande y antiguo de cuantos se conservan mundialmente, y contiene 4 petabytes de datos comprimidos en formato WARC, en los que se hallan contenidos de todos los dominios nacionales registrados en los cinco continentes –en total, más de doscientos millones de webs en más de 60 idiomas, que son ofrecidos al público con un ancho de banda de 10 Gigabits por segundo.

En cuanto a la información acerca de la infraestructura tecnológica necesaria para llevar a cabo un archivo web, la BNE, que durante 2012 siguió confiando en la Fundación Internet Archive para llevar a cabo las recolecciones con las que se engrosó la colección ya existente, vuelve a remitirse a lo estipulado con el Grupo de Patrimonio Digital en 2011: el objetivo 2 de la reunión de Vitoria se resolvería en nuestro caso aportando información relativa a lo capturado a través de dicha Fundación, ya que (aún) no contamos con nuestra propia infraestructura de captura.

Software:

- Heritrix para la captura.
- Lucene (novedad) para la indexación.
- Wayback para la búsqueda.

Hardware:

- Petabox de cuarta generación, unidad de almacenamiento y procesamiento de datos producida por Capricorn Technologies para el Internet Archive.
- Densidad: 650 TeraBytes / rack.
- Consumo: 6 KW / PetaByte.

Contenido:
>1.780 millones URLs diferentes (archivos)
85 TB

Política de colección:

archivoweb@bne.es

PASEO DE RECOLETOS, 20
28071 MADRID
TEL.: 91 580 7118
FAX: 91 5168102

La relación de los tipos de recursos web susceptibles de ser capturados para su preservación es, en el caso de la BNE, idéntica a la de 2011, en la línea de combinar recolecciones masivas y selectivas.

En la línea de combinar recolecciones masivas y selectivas, en febrero de 2012 se realizó una recolección selectiva de materias del ámbito de las Humanidades que incrementó en 44.764.995 URLs diferentes (archivos) la colección.

En el segundo trimestre de 2012 se llevó a cabo una recolección masiva del dominio .es y dominios y subdominios asociados (.com; .edu; .gob; .net; .biz; .info; .org), obteniendo un total de 301.732.926 URLs diferentes (archivos) adicionales.