Consejo de Cooperación Bibliotecaria

Grupo de Trabajo de Patrimonio digital

Consejo de Cooperación Bibliotecaria,2012
Santander, 26 y 27 marzo de 2012

Coordinadores:
Eugènia Serra. Biblioteca de Catalunya
José Luís Bueren Gómez-Acebo, Biblioteca Nacional de España

Informe del Grupo de Trabajo
Patrimonio digital

1. Miembros del grupo

Coordinadores:

Comunidad Autónoma de Cataluña
Eugenia Serra Aranda
Biblioteca de Catalunya

José Luis Bueren Gómez-Acebo
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 2011

En la reunión de Vitoria se definieron como objetivos del Grupo para el 2011:

1. Definir el tipo de recursos electrónicos web susceptibles de ser capturados para su preservación.
2. Recoger información acerca de la infraestructura tecnológica necesaria para llevar a cabo un archivo web (Catalunya, Euskadi y BNE).

3. Dinámica de trabajo del GT

No se han realizado reuniones presenciales, se ha trabajado mediante la lista de distribución de la sede colaborativa, si bien la actividad no ha sido muy intensa.

A partir de una propuesta de los coordinadores del Grupo se fijó el siguiente calendario de Trabajo:

- Objetivos 1 y 2. Recogida de información: hasta el 31 de octubre. Respecto al objetivo 2, en el caso de la BNE se trataría de tener información sobre lo que han capturado vía Internet Archive y sobre dónde se está preservando.
- Discusión y aportaciones a la información recogida sobre recursos electrónicos web: hasta el 30 de noviembre.
- Relación consensuada de recursos electrónicos web de interés para ser capturados. Hasta el 31 de diciembre.

4. Tipo de recursos electrónicos susceptibles de ser capturados para su preservación

Si bien no se ha consensuado una lista definitiva si se ha identificado por un lado la relación de recursos característicos del entorno web y si en cada caso son o no objeto de captura o depósito.

RELACIÓN PROVISIONAL DE RECURSOS SEGÚN TIPOLOGÍA

La relación de recursos es mixta, incluye tipos de documentos y de servicios. Las definiciones se han obtenido de:

- *Encyclopaedia Britannica online* http://www.britannica.com,
- *Wikipedia* http://es.wikipedia.org/wiki/Wikipedia
- ISSN http://issn.org/

TIPO DE RECURSO/SERVICIO	DEFINICIÓN	OBJETO DE CAPT/DEP
AUDIOVISUALES		SI
BLOGS	Diario online en el un individuo, grupo o organización presenta un registro de actividades, opiniones o ideas.	SI
LIBROS ELECTRÓNICOS	Archivo digital que contiene un cuerpo de texto e imágenes adecuado para ser distribuido electrónicamente y visualizado en pantalla de forma similar a la de un libro impreso.	SI
MATERIALES GRÁFICOS		SI
MEMORIAS	Publicaciones que dan cuenta de la actividad de una empresa, sociedad, institución... se publican en partes sucesivas, habitualmente con designación numérica o cronológica, pensadas para que tengan continuidad de forma indefinida.	SI
PARTITURAS DIGITALES	Versión electrónica o digital de una partitura.	SI
REVISTAS Y DIARIOS DIGITALES	Publicaciones en partes sucesivas, habitualmente con designación numérica o cronológica, pensadas para que tengan continuidad de forma indefinida.	SI

TIPO DE RECURSO/SERVICIO	DEFINICIÓN	OBJETO DE CAPT/DEP
RSS	Really Simple Syndication, anteriormente llamado RDF Site Summary o Rich Site Summary, es un formato utilizado para ofrecer a los subscriptores el contenido nuevo de sitios web que se actualizan frecuentemente.	SI, pero con dudas puesto que son contenidos muy dinámicos que no se suelen conservar mucho tiempo. Además suelen contener información que está en otros sitios web.
SITIO WEB	Colección de archivos y recursos relacionados accesibles en el World Wide Web y organizados bajo un mismo dominio.	SI
SONOROS MUSICALES		SI
SONOROS NO MUSICALES		SI
BOLETINES DE DISTRIBUCIÓN SELECTIVA	Envíos de paquetes de información a múltiples destinatarios de una red informática.	NO, excepto si se asimilan a publicaciones seriadas
LISTAS DE CORREO ELECTRÓNICO	Uso especial del correo electrónico que permite la distribución masiva de información entre múltiples usuarios de Internet a la misma vez.	NO
REDES SOCIALES (TWITTER, FACEBOOK, TUENTI, …)	Comunidades de individuos online que intercambian mensajes, comparten información, y, en algunos casos cooperan en actividades conjuntas.	NO
REPERTORIOS, BASES DE DATOS CATÁLOGOS		NO, si bien en determinados casos o materias se podría plantear la entrega de las bases de datos por parte de los editores o productores
WIKIS	Sitio Web que pueden modificar o al que pueden hacer aportaciones los usuarios.	NO, si bien en determinados casos pueden interesar por su contenido

RELACIÓN PROVISIONAL DE RECURSOS SEGÚN EDITOR O TEMÁTICA

También han comentado algunos participantes del grupo la importancia de establecer una política de gestión de la colección vinculada a la tipología en función del editor/creador del recurso o su temática, a partir de un paralelismo con los recursos tangibles. Los prioritarios son:

- Administración autonómica, provincial y local y sus organismos autónomos.
- Asociaciones, fundaciones, y otras instituciones de carácter público y/o privado.

- Webs y blogs que realicen una labor social de difusión de información y apoyo a familias (por ejemplo los de todas las asociaciones de familiares y afectados de enfermedades).
- Instituciones públicas y privadas, asociaciones,… que trabajan en la difusión y enseñanza de la lengua.
- Asociaciones e instituciones que trabajan por la Igualdad y por la Coeducación.
- Partidos políticos
- Sindicatos y asociaciones de empresarios
- Webs de deportistas individuales o de equipo de deportes minoritarios: surf, rugbi, herri kirolak,…
- Webs de Centros Tecnológicos.
- Universidades públicas y privadas.
- Empresas.
- Webs de artistas: músicos, plásticos, escritores, personalidades relevantes en la vida cultural, política, y científica.
- Blogs personales, de asociaciones, de movimientos ciudadanos, actos culturales de carácter periódico o singular (ej. Conciertos, festivales de cine, de teatro, música.
- Sitios de Congresos, Jornadas, Exposiciones… celebrados en la Comunidad Autónoma.
- Web cómic.
- Eventos sociales, deportivos, culturales… (Movimiento 15M, …)
- Elecciones
- Webs de temática de interés local.

Es necesario así mismo, tener en cuenta que la nueva ley de depósito legal durante el 2012 fijará la regulación de depósito y constitución de las publicaciones Electrónicas online lo que incide plenamente en el trabajo del GT de Patrimonio digital. Es por lo tanto imprescindible que ambos grupos se coordinen a ese respecto.

5. Información sobre la infraestructura tecnológica

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 detallados 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.

Consejo de Cooperación Bibliotecaria

RESUMEN DE DATOS DE LOS INFORMES

	PADICAT (CATALUNYA)	ONDARENET (EUSKADI)	BIBLIOTECA NACIONAL DE ESPAÑA
SOFTWARE	Heritrix para la captura	Heritrix para la captura	Heritrix para la captura
	Nutchwax, Hadoop y Wayback para la indización	Nutchwax para la indización	Nutchwax y Hadoop para la indización
	Wera y Wayback para la búsqueda	Wayback para la búsqueda	Wayback para la búsqueda
	Web Curator para metadatos y gestión	Web Curator para metadatos y gestión	
HARDWARE	4 nodos HP Proliant DL360 G4p/X3,0 con 8 GB RAM para la captura	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	PetaBox, una unidad de almacenamiento y procesamiento de datos diseñada por el Archivo de Internet y Saikley CR
	1 nodo HP Proliant DL360 G5/X5110 con 8 GB RAM para indización		
	1 nodo dHP Proliant DL360 G5/X5110 con 16 GB RAM). 1 clúster Linux de balanceo. Para búsqueda y visualización	Entorno de Producción: 3 servidores con características similares a las de un HP ProLiant DL380 G5.	

	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 86% y el otro al 77%. El resultado seria la media de los 2: 81,5%. Cuando el primer filesystem llegue al 90% las descargas programadas se empezarán a descargarse en el segundo.	
CONTENIDO	195.547 URLs	7.121 URLs	1.259.482.671 URLs
	300.000.000 archivos	22.391.057 archivos	
	9,4 TB espacio	1 TB espacio	47,5 TB
POLÍTICA DE COLECCIÓN	30.000 recursos del dominio .cat. 2 capturas/año	Diarios digitales: una descarga diaria, pero de un perfil bajo (10 MB)	2 recolecciones/año masivas de recursos dominio .es
	450 entidades con convenio. 2 capturas/año	Revistas digitales: captura según su periodicidad	
	800 webs recomendadas. 2 capturas/año	Webs: 1 captura/año	
	1000 recursos de monográficos. 1 captura/año	Blogs, captura de periodicidad variable y perfil bajo	2 recolecciones selectivas de monográficos
	30 publicaciones en serie. Captura diaria		

6. Conclusiones

Respecto a los objetivos fijados, el objetivo ***2. Recogida de información acerca de la infraestructura tecnológica necesaria para llevar a cabo un archivo web (Catalunya, Euskadi y BNE)***, se ha cumplido plenamente.

Los datos aportados por las iniciativas en curso constituyen una información valiosa que puede ser útil a cualquier otro proyecto que se quiera poner en funcionamiento. Mientras siga el trabajo del GT seria interesante actualizar dicha información anualmente.

Respecto al objetivo ***1. Definir el tipo de recursos electrónicos web susceptibles de ser capturados para su preservación,*** no se ha cumplido plenamente puesto que falta debatir y comentar más ampliamente la relación provisional que se recoge en este informe, tanto por lo que se refiere a la tipología de recursos como por procedencia y temática.

7. Objetivos 2012

- Continuar trabajando el Objetivo *1. Definir el tipo de recursos electrónicos web susceptibles de ser capturados para su preservación*

- Coordinarse con el GT de Depósito legal en la regulación relativa a Depósito Legal.

- e-Administración: 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. La Administración electrónica podría ser un buen punto de contacto, ya que es un tema de máxima actualidad en el mundo de la gestión administrativa y de los archivos

ANNEXO 1

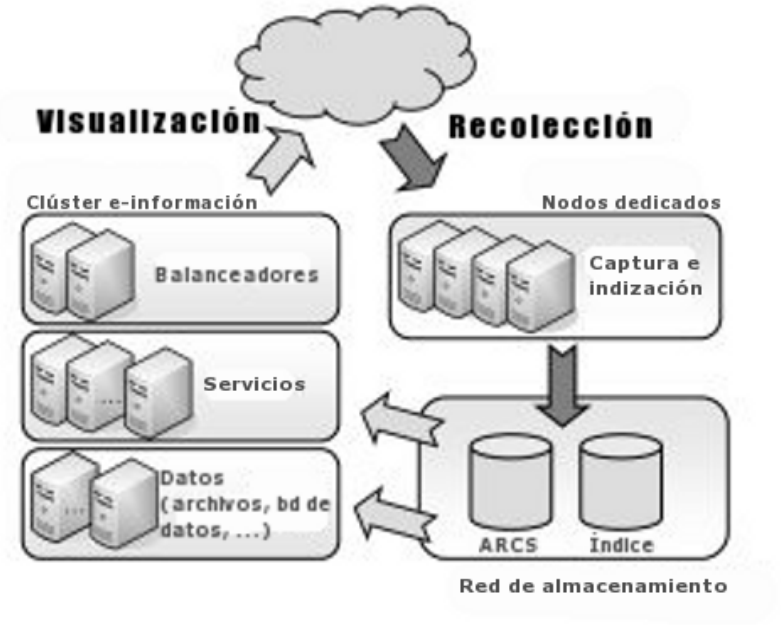
PADICAT (Patrimonio Digital de Cataluña)
Estructura técnica y capacidades (agosto 2011)

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.

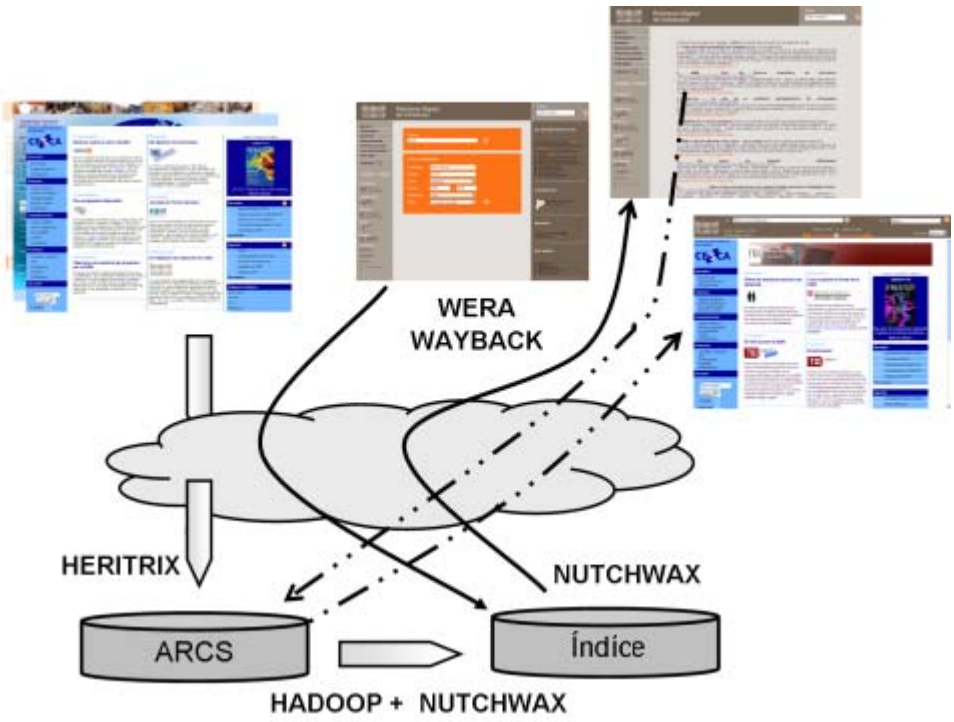


Software

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

44.374 páginas webs, identificadas por URL
195.547 capturas de versiones de las páginas web
300 millones de ficheros informáticos
9,4 TB de espacio en disco

Capacidad de crecimiento anual
75.700 versiones de 32.000 páginas web.

Política de colección

- Compilación semestral de 30.000 recursos del dominio .cat
- Compilación semestral de 550 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 86% y el otro al 77%. El resultado seria la media de los 2: 81,5%. 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.

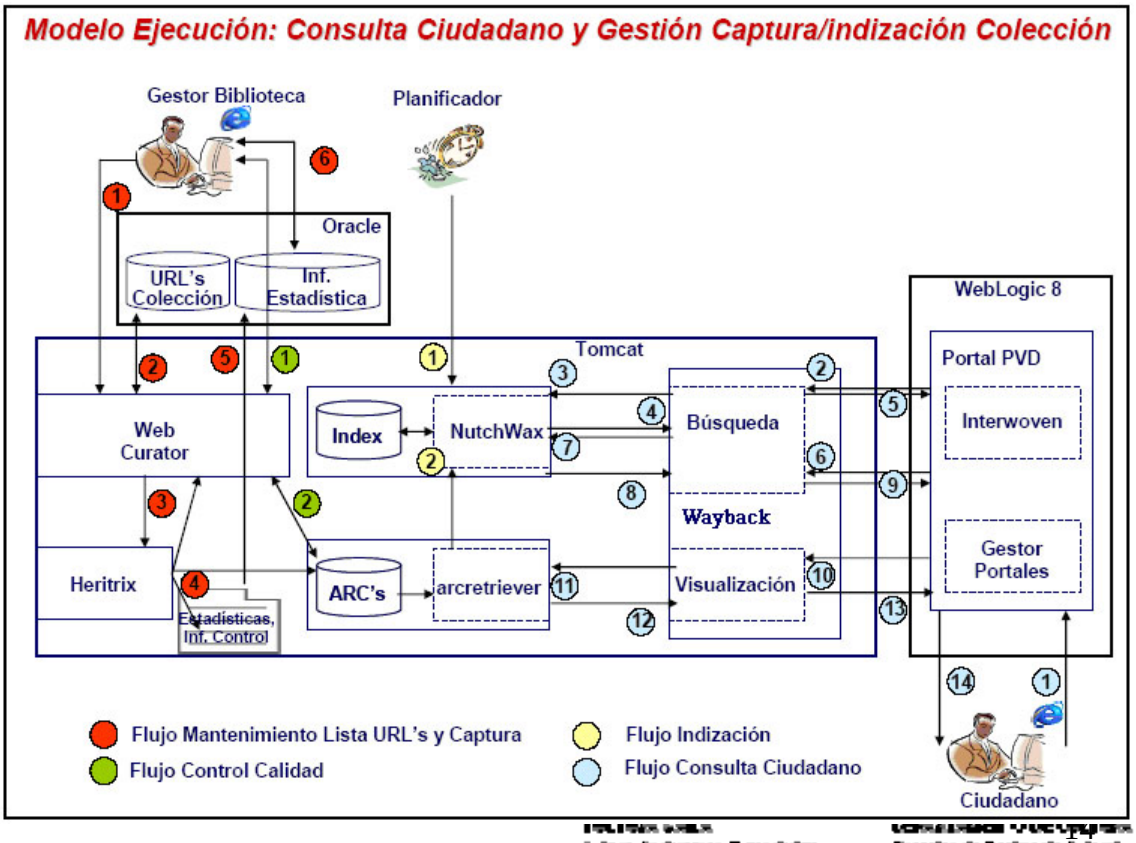
Wayback: interface de consulta del ciudadano:

- 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: una descarga diaria, pero de un perfil bajo (10 MB)
- Revistas digitales: según su periodicidad se aplican diferentes programaciones: semanal, dos veces a la semana, mensual,…

-Webs: una descarga anual. Si se observa que no ha habido variaciones se deshecha, y si se observa que ha modificado el diseño o los contenidos se captura.

-Blogs: en la actualidad muchas asociaciones y grupos optan por el blog 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 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. Debido a la constante actualización de los blogs, las capturas también varían por que pueden ser: mensuales, anuales o cada 6 meses, y un perfil bajo.

ALGUNOS DATOS

En la actualidad se tienen realizas las siguientes capturas:

- Arte (275).
- Ciencia y tecnología (52).
- Cultura (128).
- Economía y negocios (50).
- Educacion e investigación (38).
- Empresa (117).
- Euskera (102).
- Ocio y cultura (167).
- Política y gobierno (212).
- Salud (61).
- Sociedad (106).
- Sociedad de la información (68).

En cuanto a colecciones especiales que se denominan “Puntos de interés”, en Ondarenet se han realizado los siguientes grupos:

- Elecciones al Parlamento Vasco (2009).
- Montañismo vasco (2011).
- Elecciones Municipales y Forales (2011).
- Chacolí - Txakoli (2011).

En cuanto a la distribución general de la tipología de los recursos y formatos de las descargas que hemos realizado, podemos decir:

- Por descargas:
  - Por volumen de descargas: 938.167,40 MB
  - Por número de documentos: 22.391.057
- Por número de elementos:

TIPO DE RECURSOS	Nº DOCUMENTOS	%
VIDEO	3.947	0,02
AUDIO	23.759	0,11
IMÁGENES	4.246.154	18,96
TEXTO	18.117.197	80,91
TOTAL	**22.391.057**	**100**

- Por volumen de descargas (MB):

TIPO DE RECURSOS	VOLUMEN DESCARGAS (MB)	%
AUDIO	59.993	6,4 %
VIDEO	66.510	7,1 %
IMÁGENES	214.957	23,0 %
TEXTO	591.974	63,5%
TOTAL	**932.954**	**100 %**

En cuanto a las descargas realizadas hasta el momento, diciembre 2011:

- 7.121 capturas de versiones de las páginas web
- 1.376 Urls seleccionadas.

ANNEXO 3 EL ARCHIVO WEB DE LA BNE

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.
El Internet Archive es, asimismo, creador y propietario de los siguientes desarrollos de software de código abierto, que son utilizados para la recolección y archivo de la web española y forman parte del Toolkit recomendado por el IIPC:

- **Heritrix Web Crawler**, utilizado para capturar webs con calidad de archivo.
- **Wayback Machine**, empleada para reproducir y dar acceso a dicho material.
- **NutchWAX,** usada para crear índices y dar soporte a búsquedas a texto completo de archivos web, gracias a que es una herramienta idónea para la indexación de colecciones de gran volumen. El Internet Archive emplea Ubuntu Linux para su cluster de Hadoop, aunque precisa que cualquier otra distribución de Linux puede servir mientras se den ciertos requerimientos:

-Soporte x86-64/amd64 de 64-bits
-Kernel de Linux 2.6
-Sun Java 1.6, ya que tanto NutchWAX como Nutch y todas sus bibliotecas constituyentes (Hadoop, Lucene, etc.) están escritas en Java, lograr el necesario apoyo técnico es clave.

- Formato W/ARC, que permite almacenar, describir y guardar recursos primarios de la web junto con los sucesivos cambios que dichos materiales puedan experimentar a lo largo de su exposición.
- PetaBox, una unidad de almacenamiento y procesamiento de datos que fue diseñada por el personal del Archivo de Internet y Saikley CR (y que hoy día es producida por Capricorn Technologies), capaz de albergar petabytes de información (cada petabyte equivale a un millón de gigabytes).

Para el período comprendido entre 2011-12 se ha firmado un contrato por 12 meses prorrogables, con el que la BNE ha variado la prerrogativa del anterior acuerdo (2009-10), durante el que se llevaron a cabo una recolección masiva y tres recolecciones trimestrales de actualización, apostando en este nuevo tramo por una nueva fórmula mixta que combina dos recolecciones masivas y dos selectivas. La versión de Heritrix utilizada para las capturas será la v.1.3.

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

En septiembre-octubre tuvo lugar la primera recolección masiva, que duró algo más de tres semanas, tiempo requerido para recolectar los 200 millones de URLs en que se cifra el tamaño del dominio .es. Una segunda recolección masiva (deduplicación de la primera) se llevará a cabo en el segundo trimestre de 2012.
En cuanto a las selectivas, el horizonte de 6.000 seeds acordado con el IA se ha repartido como sigue: 2.000 seeds se han usado para la recolección monográfica de las elecciones del 20-N y las restantes 4.000 se emplearán para la próxima recolección selectiva, que tendrá por objeto ofrecer una muestra representativa -cultural, social y científicamente hablando- de las páginas y sitios web que componen la parcela española de Internet, prestando especial atención a la representación de redes sociales, blogs e incluso wikis y foros de discusión, entendiendo que la inmediatez informativa y el carácter efímero de que hacen gala estos recursos los hace especialmente valiosos.
El proceso íntegro de cada recolección incluye la delimitación de alcance de la misma, la ejecución de una recolección de prueba y la gestión y supervisión de la recolección propiamente dicha. Tras la finalización de cada una de las recolecciones, los ingenieros del Internet Archive llevan a cabo una revisión automatizada a modo de control de calidad tanto del contenido capturado como de los recursos web, ejecutando a continuación un rastreo adicional con vistas a capturar enlaces o imágenes que no se hubiesen archivado durante el crawl inicial.

Informes generados por el Internet Archive

Informe de la recolección: incluye el número de "seeds" (o sitios extractados de la lista inicial) recolectados, el total de documentos únicos capturados y el total de URLs recopiladas.
Informe de servidores o "host": enumera por servidor tanto el total de URLs capturadas como el tamaño total de datos capturados.
Informe de tipos MIME: detalla todos los tipos MIME recolectados y el número de objetos de cada tipo que integran la colección.
Informe de ficheros con excepciones: contiene un listado con todos los ficheros mayores de 100MB que han sido descubiertos pero no se han capturado.
Informe de ficheros excluidos: incluye una lista con aquellos ficheros que tardan más de 20 minutos en descargarse.
Una vez configurado el acceso a los servidores en los que se alojan las respectivas colecciones masivas y selectivas a través de la "Wayback Machine" del Internet Archive se procede a la indexación de lo recolectado, con el fin de poder realizar búsquedas a texto completo en la colección. El tiempo necesario para indizar los recursos recolectados es de dos semanas en el caso de las recolecciones masivas y de una para cada recolección selectiva.
La actual colección, que incluye el alojamiento de aquellos recursos web recolectados con anterioridad a 2011, reunía a principios de noviembre de este año 47.5 terabytes -a los que habría que añadir el cómputo del último tramo de la recolección del 20-N, que terminará el 23 de diciembre y recogerá dos últimos eventos clave, el discurso de investidura del nuevo Presidente y la primera reunión del nuevo Consejo de Ministros.
Existe una copia principal ("primary copy") y una copia de seguridad ("back-up copy") de todo lo archivado.

2009-10

Total de URL únicas recolectadas primera recolección masiva 2009-10 (001): 317.330.991
Total de URL únicas recolectadas en la recolección trimestral #1 (001 A) : 155.370.753
Total de URL únicas recolectadas en la recolección trimestral #2 (001 B): 156.445.148
Total de URL únicas recolectadas en la recolección trimestral #3 (001 C): 246.647.517

2011-12
Total de URL únicas recolectadas segunda recolección masiva (002 A): 94.713.560
Total de URL únicas recolectadas segunda recolección masiva 2011 A2 - 003 : 168,944,711
Selectiva 20-N: total de URL únicas recolectadas (a 6 de diciembre de 2011): 120.029.991