X Jornadas de Gestión de la Información
La dimensión del cambio: usuarios, servicios y profesionales
Biblioteca Nacional, Madrid, 20-21 de noviembre de 2008
Madrid: SEDIC, 2008, pp. 23-32

ONDARENET: EL ARCHIVO DEL PATRIMONIO DIGITAL VASCO

ONDARENET: BASQUE DIGITAL HERITAGE ARCHIVE

Pulgar Vernalte, Francisca, responsable del Servicio de Bibliotecas, Gobierno Vasco. Donostia-San Sebastián 1, Vitoria-Gasteiz, f-pulgar@ej-gv.es; **Marcos Maciá, Sonia**, Técnico Documentalista. ODEI, Pintor Gustavo de Maeztu, 4, Vitoria-Gasteiz soniam@odei.es

Resumen: El objetivo de esta comunicación es presentar el proyecto de recuperación, preservación y difusión del patrimonio digital vasco, que el Departamento de Cultura del Gobierno Vasco está desarrollando desde el 2007. Se trata de un proyecto con visión estratégica, implementado sobre herramientas de código abierto, siguiendo las recomendaciones del International Internet Preservation Consortium. A continuación se describen las principales acciones acometidas para la puesta en marcha de Ondarenet, tales como la definición y alcance del proyecto, los objetivos, la estrategia y, los medios necesarios para abordar un proyecto de esta dimensión.

Palabras clave: Repositorios digitales; Bibliotecas digitales; Archivos web, Preservación

Abstract: The aim of this writing is to present the recovery, preservation and diffusion project of the digital Basque patrimony that the Culture Department of the Basque Government has been working on since 2007. It is a project with strategic views, implemented on open code tools, as suggested by The International Internet Preservation Consortium. Next, the main steps carried out to implement the web ONDARE.NET will be described, such as the definition and scope of the project , the main goals and strategies and the necessary means to succesfully carry out such an ambitious project.

Keywords: Digital repositories; Digital libraries; Web archives; Preservation.

1. Contexto y estado del arte

El vertiginoso desarrollo de las nuevas tecnologías de la información y la comunicación permite llevar a cabo de manera relativamente sencilla tareas de digitalización de documentos impresos que, por un lado, facilitan la conservación de registros con un alto valor histórico o bibliográfico y por otro, fomentan la universalización de los contenidos, posibilitando que cualquier persona desde cualquier lugar acceda directamente a dichos documentos. Al mismo tiempo el desarrollo tecnológico y la popularización de Internet han favorecido que cada vez mayor parte de la información y el conocimiento que se genera se elaboren ya en formatos digitales tales como textos, imágenes, bases de datos, etc. y se publiquen directamente en Internet, es lo que se viene a denominar "born digital".

Nos encontramos de este modo con el denominado Patrimonio Digital al que la UNESCO define como "los recursos que son fruto del conocimiento o la expresión de los seres humanos, ya sean de carácter cultural, educativo, científico o administrativo, o comprendan información técnica, jurídica, médica o de otro tipo, y que se generan cada vez más a menudo directamente en formato digital, o se convierten a él a partir de material ya existente".

Esta nueva realidad ha obligado a que tanto los archivos como las bibliotecas trabajen en la búsqueda de nuevos modelos y estándares que permitan "adquirir, preservar y hacer accesibles el conocimiento y la información de Internet a futuras generaciones desde cualquier lugar promoviendo el cambio global y las relaciones internacionales".

Y es en este contexto en el que han surgido proyectos internacionales tales como Pandora, Minerva, Internacional Archives o Padicat, enfocados todos ellos a la recuperación y preservación del patrimonio digital. La experiencia y el conocimiento aportados por las instituciones que lideran cada uno de dichos proyectos, hacen que sean un referente y, por tanto, un modelo a seguir a la hora de poner en marcha proyectos relacionados con la preservación y difusión del patrimonio digital.

2. Ondarenet: definición y alcance del proyecto

Al hablar del universo digital de la Comunidad Autónoma de Euskadi se hace referencia tanto al conjunto de entidades e instituciones públicas y privadas productoras de elementos digitales o digitalizados, como al conjunto de elementos que componen el contenido digital y que son fundamentalmente:

- páginas web, tanto estáticas como dinámicas que contienen información de todo tipo (noticias, eventos, información cultural, etc.)
- recursos de comunicación como blogs, foros o listas de distribución.
- ficheros digitales asociados a los contenidos: documentos, imágenes, vídeos, grabaciones en diferentes formatos (.doc., .pdf., .jpg, .avi, etc.)

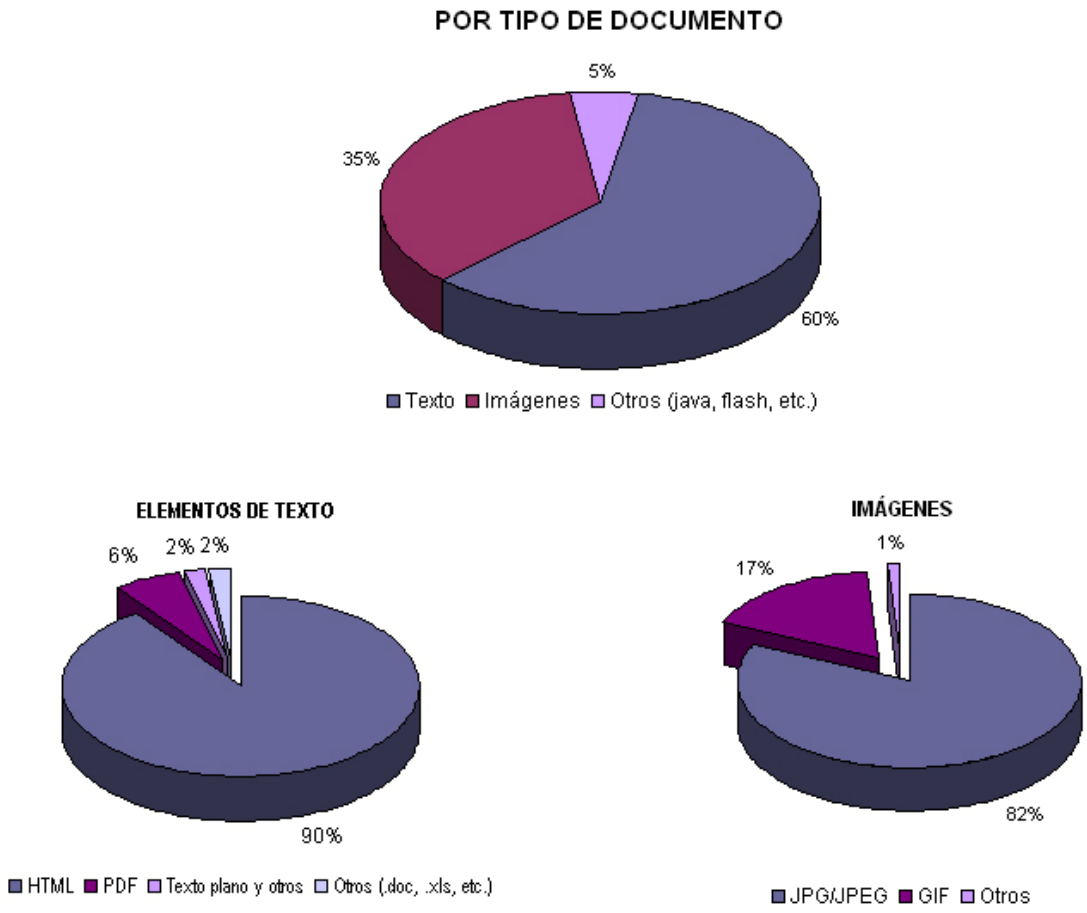

Figura 1. **Distribución del Universo Digital Vasco**

Por otro lado, la Ley 11/2007, de 26 de octubre, de Bibliotecas de Euskadi crea la Biblioteca de Euskadi, que tal y como se recoge en el punto 4 del art. 27 "se constituye en sede del patrimonio digital vasco" y entre cuyos objetivos destacan:

1. Disponer los mecanismos adecuados para garantizar la creación, preservación y difusión del patrimonio digital vasco y el acceso a él.
2. Fomentar programas de cooperación en materia de patrimonio digital.
3. Velar, en colaboración con otras instituciones, por la conservación del patrimonio bibliográfico vasco en cualquier tipo de soporte.

Para el Departamento de Cultura del Gobierno Vasco el Patrimonio Digital Vasco es "el conjunto de recursos digitales que son fruto del saber o de la expresión de la sociedad vasca en sus múltiples facetas y que por su valor deben ser conservadas para la posteridad".

Para dar respuesta a esta preocupación por conservar y preservar el patrimonio digital vasco y cumplir, al mismo tiempo, el mandato legal que la Ley 11/2007, de Bibliotecas de Euskadi establece al respecto, desde el Departamento de Cultura del Gobierno Vasco junto con la Sociedad Informática del Gobierno Vasco (EJIE) se valoró la creación de un repositorio institucional destinado a albergar los recursos digitales que conforman el patrimonio digital vasco.

Este es el inicio y punto de partida de Ondarenet, que se constituirá como el archivo electrónico del patrimonio digital vasco, siendo sus objetivos la captura, la conservación y la difusión de los objetos digitales depositados.

3. Fases del proyecto

Para el arranque de este proyecto de captura y preservación de la web vasca se elaboró una **memoria de actuación** en la que se recogen: los objetivos, la estrategia, los medios y, por supuesto, la planificación de las diferentes fases del proyecto, así como la cuantificación económica de cada una de estas fases.

3.1. Selección del soporte informático

3.1.1. Arquitectura y herramientas

En lo referente a la elección de las herramientas informáticas necesarias para llevar a cabo el proyecto se valoraron dos posibilidades: contratar un software comercial que desarrollara a medida las herramientas necesarias o utilizar el Toolkit propuesto por el International Internet Preservation Consortium (IIPC) con un desarrollo adicional llevado a cabo por una empresa especializada y supervisado en todo momento por los técnicos de EJIE. Se analizaron los pros y los contras de cada elección y se optó por la segunda de las opciones por varias razones: son las herramientas utilizadas por la mayoría de las iniciativas internacionales similares existentes, son relativamente fáciles de instalar y mantener, y son de código abierto lo que permite una total libertad en un desarrollo "ad hoc" y abarata los costes.

- **Heritrix**. El robot de captura que realiza el proceso de recolección de los componentes digitales sitios y páginas web de la colección.
- **NutchWAX**. El motor de búsqueda de código abierto que permite la búsqueda e indización de los elementos de la colección recolectados por Heritrix.
- **Web Curator**. Es la herramienta diseñada por la Biblioteca Nacional de Nueva Zelanda en colaboración con la British Library que gestiona los procesos de captura y recolección de los elementos digitales (urls) que van a componer la colección. Proporciona un interfaz web de fácil utilización a través del cual planificar y programar las capturas.
- **WERA**. Es la aplicación que hace posible al usuario final la consulta de los sitios capturados por Heritrix e indizados por WERA. Permite realizar búsquedas tanto simples como avanzadas.

Figura 2. **Herramientas del sistema de información del proyecto Ondarenet**

Una vez elegido el soporte informático sobre el que sustentar las capturas, se comenzó con la fase de implantación de las herramientas seleccionadas en los servidores de EJIE, con el fin de detectar errores y ajustar los procesos.

Durante este periodo de implantación, entre otras tareas, se optimizó la búsqueda avanzada de WERA con vistas a mejorar los resultados de las búsquedas de caracteres latinos como las tildes o la letra eñe, se automatizó el proceso de inserción de direcciones seleccionadas para las descargas simplificando de ese modo el proceso, se definió la estructura de directorios definitiva de la aplicación y se afinaron los procesos de indización.

3.1.2. Diseño del módulo de estadísticas

Asimismo se consideró como un aspecto esencial del proyecto el contar con un módulo de estadísticas que permita obtener información relevante acerca del número de descargas realizadas, de los tiempos, del volumen y tipos de los contenidos descargados, de las fechas en las que se realizan estas descargas así como de los posibles errores que pudieran haber ocurrido. La explotación de este tipo de datos posibilita ajustar las características de las descargas y optimizar las capturas.

Este módulo estadístico se realizó "ad hoc" sobre un esquema de base de datos ORACLE vinculado a una herramienta que realiza el proceso de descarga de contenidos. Esto es posible gracias a que tanto Heritrix como Web Curator Tool generan sus propios ficheros de informes (ficheros log) para cada una de las descargas realizadas de los que se puede extraer información estadística relevante.

Por último, mencionaremos que el módulo de estadísticas permite la explotación de la información de los ficheros log, mediante la emisión de diferentes informes en formato PDF, diseñados a través de la herramienta de código libre IReports lo que facilita la lectura e interpretación de los datos.

3.2. Captura y almacenamiento

3.2.1. Modelo y tipos de capturas

Entendemos por captura la descarga de un sitio web, página o componente del mismo mediante el uso de un software adecuado. Existen múltiples experiencias relacionadas con el archivo de webs nacionales que definen sus capturas según dos modelos. Por un lado encontramos el modelo integral o exhaustivo consistente en realizar una serie de "instantáneas" de la web de un país y que es el utilizado por Suecia, Austria o Noruega. Por otro lado hablamos de la captura selectiva (llevado a cabo por Australia o el Reino Unido) que consiste en realizar capturas de las web más representativas del país desde una política selectiva bien definida basada en criterios como el tema, la lengua, etc.

Ambos modelos cuentan con ventajas e inconvenientes. Mientras que una captura integral permite llevar a cabo una recolección automática a menor coste, el resultado es una colección irregular e incompleta que, por ejemplo, no accede a recursos de la denominada "Internet invisible". Por otro lado, el modelo selectivo permite la creación de colecciones equilibradas pero suponen un alto coste y pueden resultar un tanto parciales. Con el tiempo ambos modelos han dado paso a uno nuevo denominado híbrido que combina la captura sistemática de la web nacional con acuerdos con instituciones productoras según los intereses temáticos.

Con el fin de conseguir los objetivos marcados en el proyecto de patrimonio digital vasco, se ha optado por seguir un modelo híbrido que aglutina tanto procesos de recolección integral regional consistente en la realización de instantáneas de la web vasca en Internet de forma periódica, como de recolección selectiva y temática basada en la captura de urls previamente seleccionadas y que sean de interés tanto por sus contenidos, como por las características del productor. La captura integral se lleva a cabo sobre una serie de sitios web completos que en su conjunto componen una imagen representativa de la web vasca bien por estar albergados en el País Vasco,

pertenecer a entidades relacionadas con Euskadi o estar en euskera. La captura selectiva se corresponde a urls de interés por su temática y alcance que requieren de un mantenimiento manual. Se prevé realizar dentro de este apartado las denominadas "capturas sobre eventos y hechos relevantes" (elecciones, exposiciones, etc.) con el fin de conformar colecciones especializadas.

3.3. Elaboración de un sistema de clasificación

Con el fin de facilitar la búsqueda y recuperación de la información era necesario contar con un sistema de clasificación por materias que permitiera indizar las webs descargadas de manera unificada facilitando la localización de los recursos capturados a través de un sencillo índice de navegación.

Se consultaron las clasificaciones utilizadas por proyectos similares como Padicat, Pandora y UK Web Archive y se comprobó que todos estos proyectos utilizan un número más o menos reducido de grandes grupos de temas, subdivididos, a su vez, en un segundo nivel de materias más específicas. Siguiendo ese mismo esquema se elaboró una clasificación propia, dividida en 12 temas principales:

Arte	Euskera
Ciencia y tecnología	Ocio y cultura
Cultura Vasca	Política y gobierno
Economía y negocios	Salud
Empresa	Sociedad
Educación e investigación	Sociedad de la información

3.4. La difusión: la interfaz de consulta

Aunque Ondarenet se configura como el archivo electrónico del patrimonio digital vasco, su finalidad prioritaria es facilitar su acceso y consulta a los usuarios. Con este fin se ha diseñado una interfaz de consulta que permita de una forma sencilla, amigable e intuitiva la localización de los recursos capturados. Esta interfaz se encuentra integrada en la página web del Servicio de Bibliotecas del Gobierno Vasco, accesible a través de la dirección (http://www.kultura.ejgv.euskadi.net/r46-4878/es), y se estructura en tres tipos de consultas:

- *Búsqueda simple*, permite recuperar la información bien introduciendo el término o términos de búsqueda en el cajetín "Texto" de forma que cuando el usuario lanza una búsqueda a través del interfaz lo hace de forma similar a como se realiza con un buscador tradicional. Asimismo es posible teclear la dirección completa en el cajetín "Url" si lo que interesa es recuperar las capturas de una url concreta.
- *Búsqueda avanzada*, permite delimitar el término o términos de búsqueda mediante parámetros como el formato (imagen, sonido, etc), la fecha y la colección. Es posible, además, especificar el orden en el que se quiere recuperar los resultados.
- *Índices*, permite realizar búsquedas a través de un índice basado en la clasificación utilizada para indizar las webs capturadas.

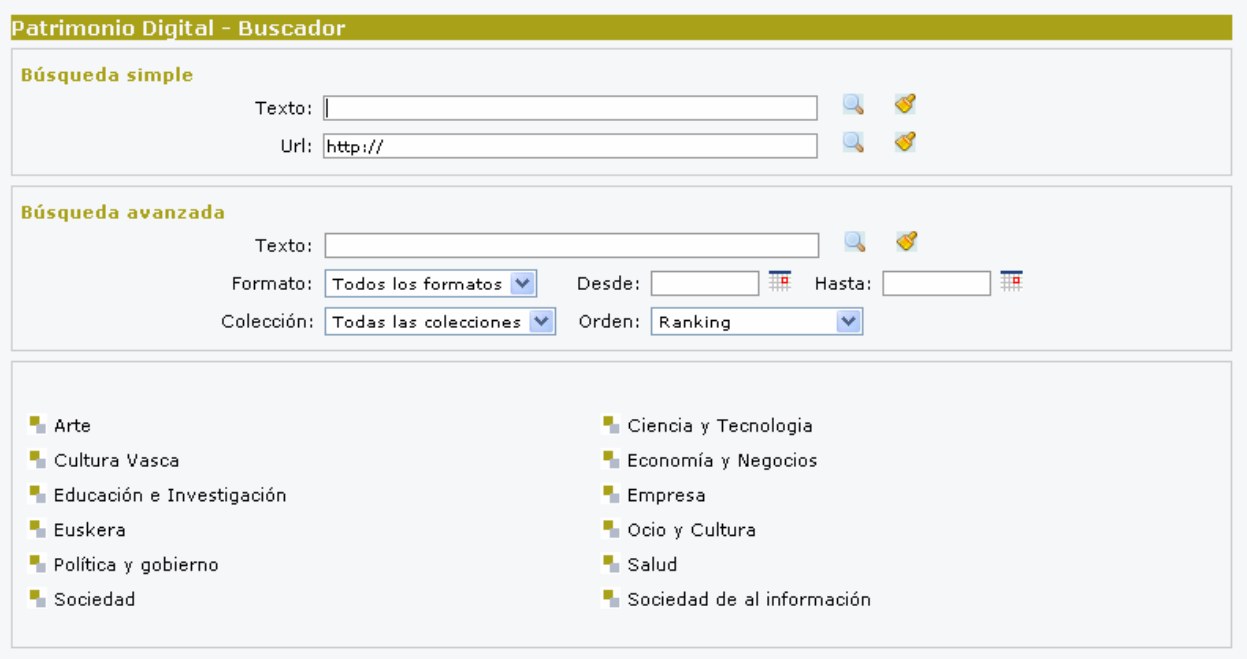

Figura 3. **Interfaz de búsqueda de Ondarenet**

Las búsquedas realizadas devuelven los resultados de manera similar a como los presentan buscadores como Google o Yahoo resaltando el término o términos buscados entre el contenido de la url capturada.

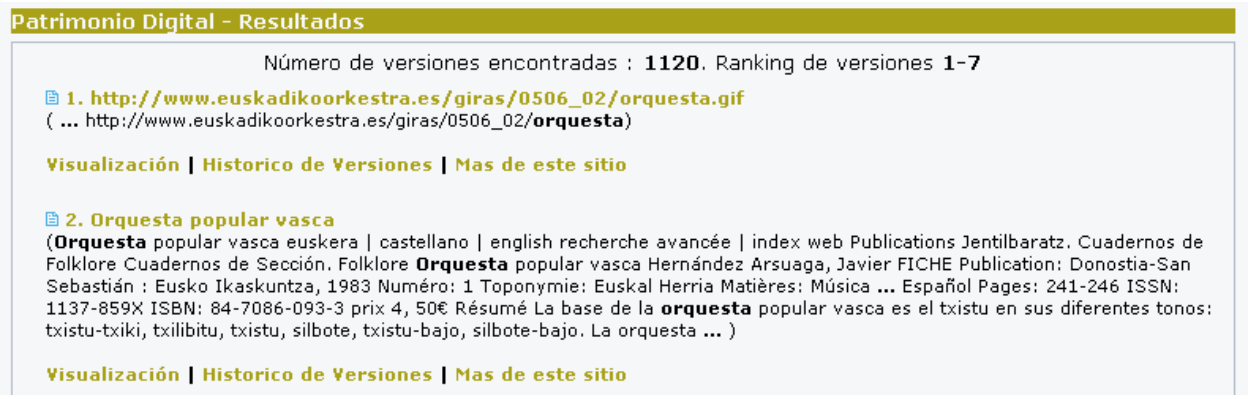

Figura 4. **Resultado de búsqueda en Ondarenet**

4. El proceso de implantación

4.1. Primeras pruebas

Una vez implementado el soporte informático y seleccionado el modelo de captura se realizaron una serie de pruebas de capturas reales con el fin de detectar errores, estimar los tiempos de descarga, etc. Se decidió iniciar la captura con una relación de webs pertenecientes a cada uno de los 12 grandes temas de la clasificación elaborada

para lo que se seleccionaron una serie de webs representativas de cada una de ellas. El análisis de estas capturas de prueba mostró que algunos de los sitios seleccionados se descargaban de forma errónea o incompleta por lo que se realizó una selección más amplia con el fin de ir descartando los errores detectados. Finalmente se consiguió la captura correcta de 15 sitios webs.

El volumen de descargas en el entorno de pruebas de estos primeros 15 sitios web fue de 9.5Gb, y el tiempo medio de descarga de 2 horas y 49 minutos. En el siguiente cuadro se refleja el volumen y tiempo de descarga de cada una de las webs seleccionadas.

Grupo	**Subgrupo**	**Target**	**Duración (dd/hh/mm/ss)**	**Volumen**
Ciencia y tecnología	Investigación y desarrollo	http://www.inguma.org/	00: 05:24:17	3.1 Gb
Cultura vasca	Etnografía y folklore	http://www.eusko-ikaskuntza.org	01:04:06:16	2.56 Gb
Política y gobierno	Administración local	http://www.eudel.net/	00:00:27:50	823.76 Mb
Salud	Asociaciones y fundaciones	http://www.bioef.org/	00:02:02:35	765.69 Mb
Educación e investigación	Formación no-reglada	http://www.isei-ivei.net/	00:01:21:21	726.94 Mb
Cultura Vasca	Historia	http://www.berrikuntza.net	00:00:54:03	512.99 Mb
Educación e investigación	Formación no-reglada	http://www.berritzeguneak.net/	00:02:24:32	323.25 Mb
Política y gobierno	Administración local	http://www.zeberio.net	00:00:19:43	223.52 Mb
Ocio y cultura	Archivos, bibliotecas y centros de documentación	http://www.eresbil.com	00:01:10:50	94.24 Mb
Arte	Musica	http://www.euskadikoorkestra.es/	00:00:57:06	90 Mb
Empresa	Asociaciones y fundaciones	http://www.euskolabel.net/	00:00:54:43	68.45 Mb
Sociedad	Religión	http://www.santuariodeloyola.com/	00:00:08:51	15.06 Mb
Euskera	Asociaciones y fundaciones	http://www.bagera.net/	00:00:09:04	13.92 Mb
Economía y negocios	Asociaciones y fundaciones	http://www.eke-fce.com/	00:00:05:33	8.53 Mb
Sociedad de la información	Portales temáticos	http://www.jalgi.com/	00:00:01:07	981.41 Kb

Observamos la gran diferencia que existe entre los tiempos de descarga de las dos urls con mayor volumen de información: http://www.inguma.org/ (3.1 Gb) y http://www.eusko-ikaskuntza.org (2.56 Gb). Mientras que la primera se descargó en cerca de 5 horas y media, la segunda necesitó más de un día para su descarga completa. Por ello, y dado que los tiempos dependen del estado de la red y de los agentes de descarga, parece deducirse que estas mediciones no resultan del todo representativas en los casos de sitios con mayor volumen de información.

5. Ondarenet: balance y perspectivas

Esta comunicación viene a ser un resumen del intenso trabajo que a lo largo de un año ha llevado a cabo un equipo formado por técnicos del Departamento de Cultura y de EJIE.

Es evidente el largo camino recorrido desde la definición y puesta en marcha de un proyecto de tal magnitud, sobre todo si se tiene en cuenta que la memoria de actuación fue aprobada en octubre de 2007. A pesar de valorar positivamente los resultados obtenidos, somos conscientes de que estamos sólo al inicio de un largo camino y que, evidentemente, surgirán nuevos retos a los que deberemos hacer frente.

5.1. Modelo de descripción

La selección y captura de los sitios web es tan sólo la primera parte del proyecto. Desde el Departamento de Cultura del Gobierno Vasco se ha establecido, además, como un objetivo prioritario la descripción de los recursos capturados por medio de estándares internacionales que permitan una descripción completa y una posterior recuperación de la información, garantizando a su vez la interoperabilidad con otros sistemas.

Web Curator Tool, posibilita la programación de las descargas de los sitios web, y permite describir los recursos capturados mediante el estándar Dublin Core. De esta manera es posible añadir campos como título, autor o materia mediante un breve formulario compuesto por los 15 principales campos recogidos por Dublin Core Metadata Initiative. A pesar de ello, es importante mencionar que no se trata de una herramienta de descripción y dichos datos no son recuperables, sino que la información se almacena en ficheros ARC, un formato que sólo puede ser leído por Nutch Wax.

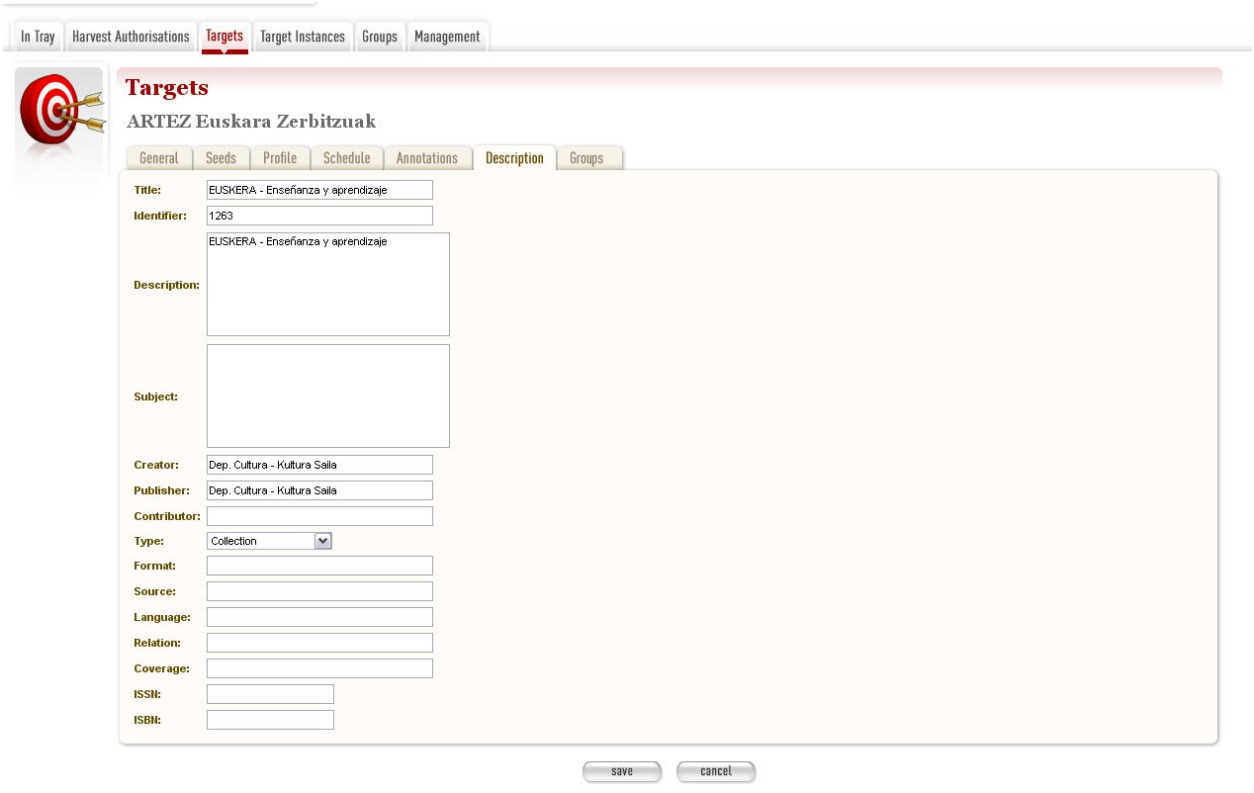

Figura 5. **Ficha descriptiva en Web Curator**

Así, uno de los retos de futuro del proyecto es conseguir que los resultados de las búsquedas se presenten en fichas descriptivas basadas en esquemas de datos XML para estándares de descripción internacionales (Dublin Core, MODS, METS, etc.), y de esa forma convertir Ondarenet en un repositorio institucional que cumpla con el protocolo OAI-PMH para la comunicación e intercambio de metadatos.

Referencias bibliográficas

Cócera, Daniel; Llueca, Ciro. "PADICAT: realitat i reptes de 3 anys d'arxiu web de Catalunya". En: *Jornades Catalanes d'Informació i Documentació*, 2008, pp. 163-178.

Llueca, Ciro. "Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales". En: *BiD: textos universitaris de biblioteconomia i documentació* 2005 desembre, n.15. Consultado en: 12-08-2008. *http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec2.htm*

Paynter, Gordon; Joe, Susana; Lala, Vanita; Lee, Gillian. "A Year of Selective Web Archiving with the Web Curator at the National Library of New Zealand" .En: *D-Lib Magazine*, 2008 May/June, v. 14, n. 5/6. Consultado en: 12-08-2008. *http://www.dlib.org/dlib/may08/paynter/05paynter.html*

Plan Vasco de la Cultura. Vitoria-Gasteiz : Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia = Servicio Central de Publicaciones del Gobierno Vasco, 2004. ISBN 84-457-2166-6.

Serra, Eugènia. "Archivando la Web catalana: iniciativas cooperativas de preservación digital en Catalunya". En: *La Recuperación de la memoria, muchas más oportunidades que realidades: el trabajo cooperativo de archivos, bibliotecas y museos*. Universidad del País Vasco, 23-25 2006. Consultado en: 12-08-2008. *http://www.bnc.es/bc/archivando_web_catalana.pdf*

UNESCO. Directrices para la preservación del patrimonio digital, 2003. Consultado en: 09-08-2008. *http://unesdoc.unesco.org/images/0013/001300/130071s.pdf*