CAPTURING THE BASQUE WEB

Francisca Pulgar Vernalte, Head of the Library Service, Basque Government, e-mail: f-pulgar@ej-gv.es; Sonia Marcos Maciá, Documentalist. ODEI, e-mail: soniam@odei.es

Launched in 2007 by the Basque Government Department of Culture and the Basque Government IT services company, the Basque web capture project joins forces with European and international initiatives to harvest and conserve the online digital heritage. Projects launched in the mid-90s both by national libraries and private organisations and which have given an enormous boost to the development of tools and directives for conservation of the born digital heritage.

In this document, we describe the setting in motion of Ondarenet (www.euskadi.net/ondarenet), an information system permitting collection, conservation and distribution of the Basque web. The software used for this project is the Toolkit proposed by the International Internet Preservation Consortium (IIPC). Used by most existing similar international initiatives, these open-source tools are relatively easy to install and maintain (Heritrix, NutchWAX, WebCurator and Wayback).

This is an ambitious project which, like all historical archives, harnesses documents and creates collections for their custody with a view to the preservation of a country's historical memory, a memory in this case presented in the shape of collections consisting of born-digital resources.

Key words: web archive, digital heritage, digital preservation

1. Digital heritage

There's nothing new in stating that the generalised use of information technologies, and particularly of the Internet, has caused a turning point in the way information is produced, distributed and accessed. Low-cost, easy-to-use digital publishing tools have brought about an explosion of born-digital documents, changing the way information is produced at lightening speed. Here we have to quote an important figure; in the Spain of 2007, 75,600 titles were published on paper and 7,503

in electronic format. However, 228,097 web sites with the .es domain were registered. In other words, the content published in digital format accessible over the Internet almost tripled that published in traditional formats.

We therefore find ourselves faced with a new situation of having to develop new models and standards of harvesting, conserving and distributing the enormous amount of born-digital information, the entirety of which unquestionably constitutes what is known as a country's digital heritage. The history of a country's knowledge would be incomplete without the addition of electronic web resources to library collections.

These resources contain essential information for teaching future generations about a country's social, cultural and political habits; information that the memory institutions have the obligation to conserve and distribute, just as they have been doing to date with material printed in traditional formats and delivered by legal deposit to the national libraries.

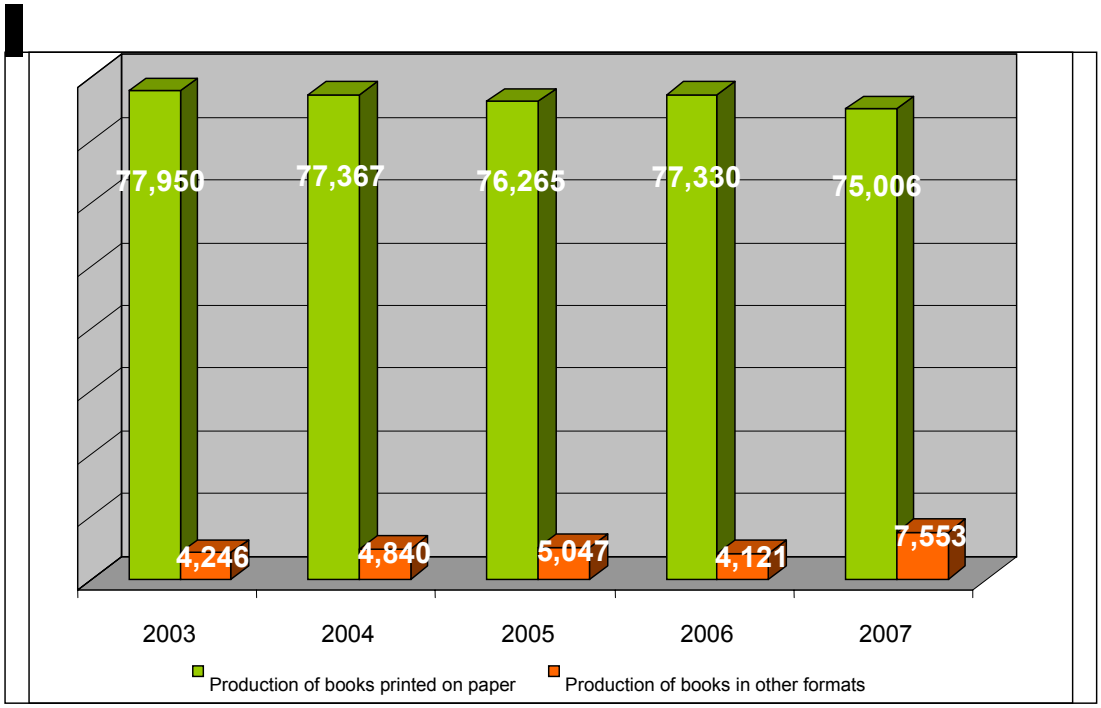

Fig. 1. Evolution of printing in Spain by format

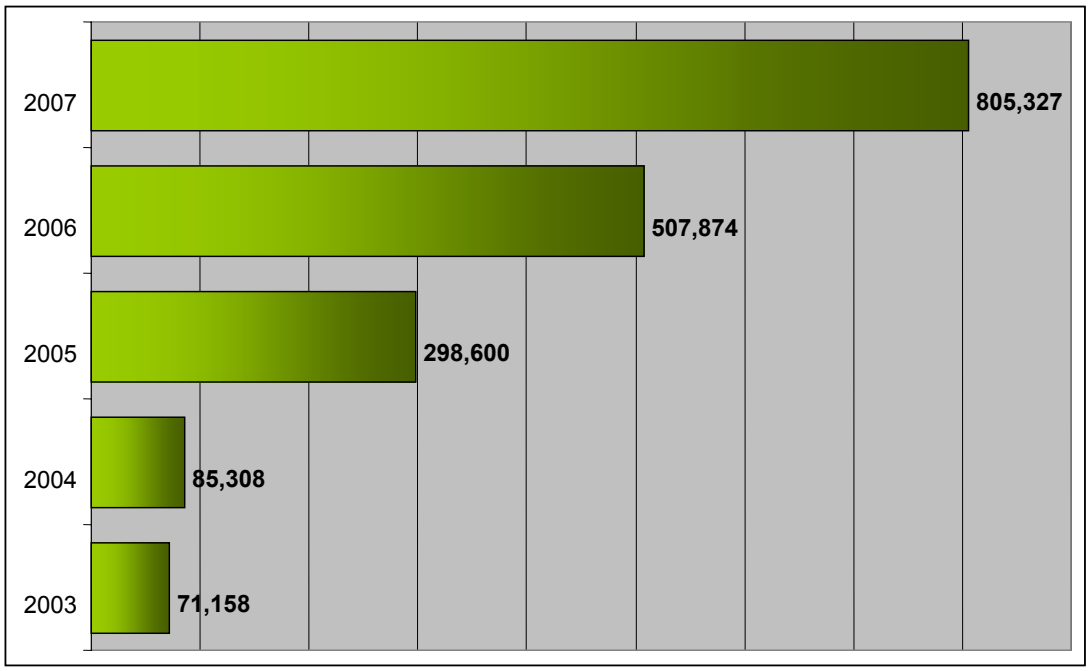

Fig. 2. Evolution of the number of .es domain names registered in Spain (2003-2007)

2. Archiving the web: origins, initiatives and access to information

Ever since the mid-90s, when the Internet definitively installed itself as a key information resource, numerous organisations have been studying the need to archive and preserve information characterised by immediacy, a short life-span and change.

The pioneer and most famous project in this respect is the Internet Archive, a not-for-profit organisation founded in 1996 to collect and conserve Internet resources and preserve the web memory. The project flourished thanks to the support of organisations like Alexa Internet or the Library of Congress and, since 2001, offers access to its collections thanks to a tool known as the “Wayback Machine”, a database permitting users to harness information from among the millions of pages stored in the different collections of web pages, texts, images or videos.

Parallel to the above, the national libraries in different countries also started launching different experiences in 1996, all targeting the selection and capture of what could be called the “national web heritage”, resulting in the development of different models of capture.

In the first place is the bulk or exhaustive model created to take snapshots of the entire web based on criteria like language, domain or server location. Outstanding among the initiatives using this comprehensive model is Kulturarw3, spearheaded by the National Library of Sweden since 1996 and which can be considered as the first endeavour by a national library to research the use of harvesters to capture and archive information available on the web. Although the earlier versions were limited to capturing web sites with the .se domain, searches have now been extended to include other domains (.nu), web sites lodged in Swedish servers and web pages written in Swedish.

There is also a selective model, the purpose of which is to harvest web sites based on a restrictive policy focussing on the quality of the resources, their subject, or their direct relationship with the national reality. The main exponent of this model is Pandora (Preserving and Accessing Networked Documentary Resources of Australia), headed by the National Library of Australia, which was another of the first national libraries to approach a web archiving project. This initiative concentrates on the selection of online publications and websites about Australia, written by Australian authors or about Australian subjects. The capture is implemented based on its own software, Pandas, which is also used in other projects. One essential aspect of this project is that the access to the harvested resources is backed by the authorisation to distribution rights negotiated with each publisher.

Both models have their pros and cons. While opting for bulk harvesting as a method of automatic harnessing is cheaper, the result is an incomplete collection given its failure to access resources such as those found in the invisible Internet. On the other hand, although using the selective model is more expensive, it permits more balanced, if partial collections.

Lastly, we have the hybrid version born from a fusion between the two above-mentioned models which complements the periodical web snapshot with selective actions. One example of this model is the netarchive.dk initiative led by the Danish Det

Kongelige Bibliotek since 1998. In 2005, introduction of the new Legal Deposit law meant that the Danish library was able to obtain a complete overview of the Danish domain by taking a snapshot of all websites with the .dk domain four times a year. Since then, it has employed the hybrid system with a triple objective: to take a snapshot of the .dk domain, to harvest frequently used web pages –not reflected by the snapshot- and to harness web pages dedicated to an event and which are destined to disappear once the event in question is over.

Another form of archive, which we could place in the selective model category, is that related to collecting and preserving web content referring to a specific subject or event. In this respect, in 2000, the Library of Congress launched MINERVA (Mapping the Internet Electronic Resources Virtual Archive), having already created more than 17 web collections related to subjects including the 2000, 2002, 2004 and 2006 elections, the Iraq war or events surrounding September 11th.

Taking as a reference the magnificent PADI: gateway to international digital preservation resources and ICADS (IFLA-CDNL Alliance for Digital Strategies), 19 countries now have web archive projects at different stages of introduction. In Spain, curiously, the first web capturing experience wasn't spearheaded by the Biblioteca Nacional de España, but by the Biblioteca Nacional de Catalunya, which, in 2005, launched the first Spanish web archive project, Padicat, the aim of which is to "harvest, process and provide permanent access to all Catalan cultural, scientific and general digital format productions", the experience and collaboration of which served as a reference for the setting in motion of Ondarenet.

One essential part of archiving information is its distribution. In this respect the number of national web deposits permitting free access to their collections is extremely limited. Often the impediment is the possibility of conflicts arising from the distribution of resources captured with no express authorisation, given that most countries have no clear legislation in this respect.

3. ONDARENET: the Basque web archive

The starting point for the creation of Ondarenet is to be found in the Basque Plan for Culture approved by the Basque Government in 2004 with the idea of fostering Basque culture in and beyond the Basque Country.

> "….Cultural heritage refers to a country's historical memory. It can be immaterial or material, immovable or movable. Its fundamental objectives are the preservation, conservation, restoration, revaluation and dissemination of a community's cultural legacy. It is normally managed by cultural heritage centres or services (ethnographic, architectural, archaeological, artistic, industrial...), the museums system, the archives system and the libraries system. It consists of public and private ownership structures, with a clear vocation to serve as a public service, receiving strong institutional backing, and is almost always a non-profit-making endeavour…" (Basque Plan for Culture).

The existence of an own language has similarly given shape to an own culture differing from those existing in the rest of Spain and expressed in activities, creations, traditions, beliefs or events from the past linked to important forms of expression of the culture and ways of life of the Basque people. Thus, Law 7/1990, of 3 July, on the Basque Cultural Heritage, defines the Basque Cultural Heritage as the principal expression of the identity of the Basque people and the most important witness to the historical contribution of this people to universal culture.

The Basque Country covers a total area of 7,234 km² and has a population currently standing at 2,155,546 inhabitants. This is a community with two official languages: Basque and Spanish. Unlike the other modern Spanish languages, Basque neither comes from Latin nor belongs to the Indo-European family and is considered as being possibly the oldest language on the European continent. Given this singularity, the debate on the origins of the Basque language has a long history in which European historians and linguists from different periods and countries have taken part. Since

1979, the year in which the Basque Statute of Autonomy was approved, Basque has been, alongside Spanish, the joint official language of the Basque Autonomous Community.

If the Basque Plan for Culture served to foster consolidation and progress of the actions taken with respect to cultural heritage, the publication of Law 11/2007, of 26 October, on Basque Libraries, opens a new prospect. This law effectively created the Biblioteca de Euskadi or Basque Library, an institution responsible for collecting and custodying the Basque bibliographic heritage and, of course, its digital heritage counterpart.

The Basque Library, converted by law into the Basque digital heritage headquarters, has the mission of launching programs promoting the safekeeping of digital heritage, hence the setting in motion of Ondarenet, the title of the project described herein, an acronym corresponding to the Basque name for the electronic digital heritage archive and constituting an information system essentially aimed at collecting, conserving and distributing the Basque digital heritage.

4. Gestation of the project

The Ondarenet project was launched in 2007 by the Basque Government Department of Culture with the backing of the Basque Government IT services company (EJIE). It started by composing a road map to define the objectives, tasks and scope required of the online Basque web archive.

4.1. Choice of computer tools

We considered two different computer tool options: the hiring of commercial software for customised development of the necessary tools or using the Toolkit proposed by the International Internet Preservation Consortium (IIPC), with an additional development provided by a specialised company and constantly supervised by the EJIE technicians. We analysed the pros and cons of each choice, finally opting for the second of the two for various reasons: these are the tools used by most existing similar international initiatives, they are relatively simple to install and maintain, and the

fact that they are open-source permits complete freedom in an ad-hoc development while reducing costs.

1. Heritrix. The web crawler designed to harvest the digital components of the web sites and pages in the collection.
2. NutchWAX. The open-source web archive collection search engine for searching and indexing the web sites and pages harvested by Heritrix.
3. Web Curator. The tool designed by the National Library of New Zealand in collaboration with the British Library to manage the capturing and harvesting of digital material (URLs) making up the collection. This tool comes with an easy-to-use web interface for planning and scheduling captures.
4. WERA. The application permitting end-users to consult the sites harvested by Heritrix and indexed by WERA. Permits both simple and advanced searches.
5. Wayback. This tool is used to index and subsequently view downloaded content using Web Curator. The tool itself is capable of reading and understanding WCT-generated content; generating an index structure with the information captured and providing visual access to the information in question (by means of a web interface); regenerating the original page for this purpose.

4.2. Definition of a selection strategy

By digital universe in the Basque Autonomous Community, we refer to all public and private entities and institutions producing digital or digitized material, and to all elements making up the digital content. These are basically:

- Web pages, whether static or dynamic, containing all kinds of information (news, events, cultural information, etc.).
- Communication tools like blogs, forums or distribution lists;

- Associated content files: documents, images, videos, recordings in different formats (.doc., .pdf., .jpg, .avi, etc.)

As explained in point 2 of this document, there are three major models of capture: bulk, selective and hybrid. In the case of the Basque Country, we are faced with an important hurdle when it comes to bulk harvesting: the non-existence of an own domain directly linked to the online Basque linguistic and cultural community. Given the situation, we decided to go about a selective, subject-related harvest based on the capture of previously selected URLs of interest due both to their content and to the characteristics of their producers. We therefore drew up a list of almost 600 URLs corresponding to web sites related to culture, economy, politics or science created in Basque or maintained by Basque organisations. We also planned so-called "captures on important events and happenings" with a view to compiling specialised collections preserving the web pages created for the purpose of current events.

4.3. Devising the classification system

To streamline the information search and recovery process we needed a subject classification system permitting unified indexing of the downloaded web sites for easy location of resources captured based on a simple browsing index.

Consultation of the classifications used by similar projects like Padicat, Pandora and the UK Web Archive showed that they all use a fairly limited number of major subject groups, which are in turn sub-divided into a second level of more specific subjects. We observed the same outline in creating our own classification, which we have divided into 12 major subjects: art, Basque language, science and technology, leisure and culture, Basque culture, politics and government, economy and business, health, society, education and research and information society. Each of these subjects is in turn sub-divided into more specific subjects, meaning that in leisure and culture, for example, we find the subdivisions: archives, libraries and documentation centres.

4.4. Defining a distribution strategy

Although Ondarenet is configured as the electronic archive for the Basque digital heritage, its priority purpose is to facilitate access and consultation by users. Thus, with a view to distributing the captures made, a subsite has been designed within the Department of Culture web page (www.euskadi.net/ondarenet), on the one hand permitting simple searches of the material archived, and on the other providing users with an e-mail address for submitting recommended captures.

4.5. The technical team

One important aspect was to establish the professional skills and number of staff necessary to ensure that the project was properly implemented. A project manager was appointed to plan the tasks required and coordinate the technical team. Implementation and backup for the IT tools fell upon the EJIE technicians with the support of three external technicians, all familiar with J2EE architecture. For the purposes of documentary management, the project has a head librarian (Head of the Library Service) who is in charge of coordinating the documentary and administrative aspects of the project, and a documentalist responsible for updating the capture lists and descriptions of the items in the collection.

5. Launching the system

The IT implementation of the project was expensive due to requiring the installation of new products not corresponding to the standard versions used by the Basque Government. It was decided to split the launching process into two stages: pilot and global, in order that the former would make it possible to obtain specific details on volumes, times and problems associated to downloading and permit the introduction of corrective measures based on the experience gained.

Hence, for the pilot test running from September to November 2008, a list of web sites falling under one of the 12 main classification headings was selected. Following a series of tests and wrong or partial downloads, we were finally able to

correctly harvest 15 web sites corresponding to a download volume of 9.5Gb with an average download time per resource of 2 hours and 49 minutes.

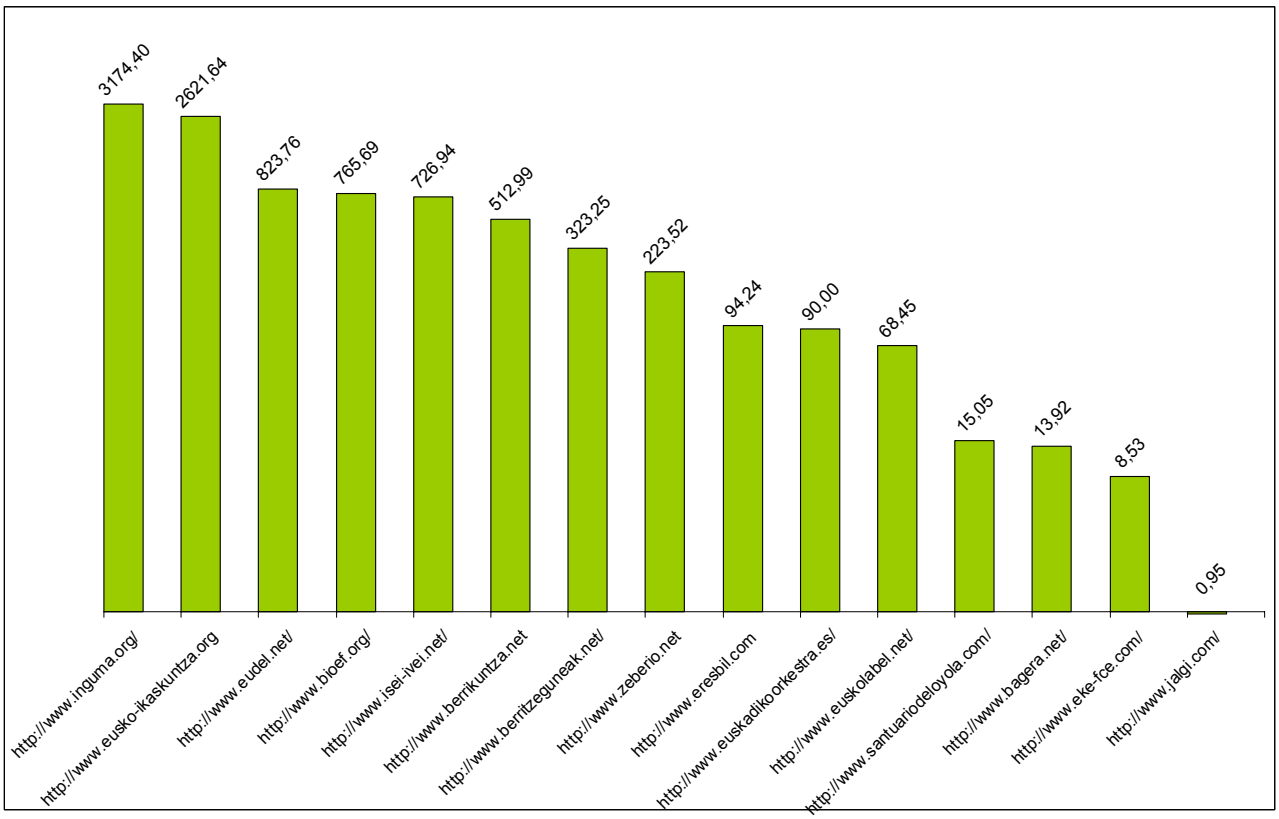

Fig. 3. Volume of the first captures (in Mb)

Later, coinciding with the Basque autonomous elections, we organised the capture of the webs of political parties, trade unions, candidates' blogs, etc., a collection grouped under the new heading "Basque Parliament elections 2009", given that the elections are unquestionably a relevant social occurrence of an importance and significance deserving study and analysis in the near future.

6. ONDARENET: online Basque digital heritage

Undeniably, the most costly part of launching a project of this size is the amount of preparation required from the moment of drawing up the report on action to be taken until introduction of the necessary software, not forgetting, of course, the period

required to test the product and ensure that the desired objectives are being achieved. Despite positive assessment of the results obtained, we realise that we have only taken our first steps on a very long road and that new challenges requiring response will obviously appear along the way.

The Basque bibliographic heritage, initially consisting of collections published by traditional methods, on paper for instance, is gaining in wealth and scope with the tremendous amount of resources either digitized or directly born-digital.

With Ondarenet we are now capturing webs related to the Basque language, culture and society. We have created a URL archive including political bodies, associations, universities, blogs, etc., all representative of the Basque society. The idea is to make two captures a year for each web site.

Thanks to this initiative, we will shortly be able to start offering citizens an interesting collection of digital resources essential for studying and gaining knowledge of the Basque Country.

However, in the near future we will also proceed to capture and conserve publications directly published on the Internet. This digital collection on the history, culture, politics, etc. of the Basque Country will be completed with works from the Basque bibliographic heritage, digitized and archived in the Basque digital library (www.euskadi.net/liburutegidigitala).

We have also already taken the necessary steps to convert Ondarenet into an institutional repository destined to house the resources making up the Basque digital heritage. This repository will comply with the OAI-PMH protocol for communicating and exchanging metadata, and will house all written digital resources. This initiative will therefore satisfy the concerns of the Department of Culture with regard to conserving and preserving the Basque digital heritage, while complying with the legal mandate established in this respect by Law 11/2007 on Basque Libraries.

Bibliography

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

Dalbello, Marija (2008) Circulating culture for the knowledge continuum: living history, digital history and the history web . In *Pacevicius, Arvydas and Manzuch, Zinaida*, Eds. Proceedings Memory in Digits: Communication of Memory in Archives, Libraries and Museums: The Interaction of Science, Policy and Practice, pp. 34-47, Vilnius University. Retrieved april 2009 from http://dlist.sir.arizona.edu/2477/01/Dalbello%5FMID%5F2008.pdf

Day, M. (2003). Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust. . Retrieved april 2009 from http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

España. Ministerio de Cultura. Panorámica de la Edición en España. Retrieved april 2009 from http://www.mcu.es/libro/MC/PEE/estadisticas/globalesEvo.html

España. Ministerio de Industria, Turismo y Comercio, Dominios.es. Retrieved april 2009 https://www.nic.es/index.action

Llueca, C. (2005). Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales. In *BiD: textos universitaris de biblioteconomia i documentació* (15). Retrieved april 2009 from http://www2.ub.edu/bid/consulta_articulos.php?fichero=15lluec2.htm

Llueca, C. (2008). El archivo de Internet: la experiencia del proyecto PADICAT (Patrimonio Digital de Cataluña. In *IV Encuentros de Centros de Documentación de Arte Contemporáneo*, Vitoria-Gasteiz (Spain)

National Library of Australia. PADI, Preserving Access to Digital Information. Retrieved april 2009 from http://www.nla.gov.au/padi

Paynter, G., Joe, S., Lala, V. & Lee, G. (2008). A Year of Selective Web Archiving with the Web Curator at the National Library of New Zealand In *D-Lib Magazine*, 14

(5/6) Retrieved april 2009 from http://www.dlib.org/dlib/may08/paynter/05paynter.html

Plan Vasco de la Cultura. (2004) Vitoria-Gasteiz : Central de Publicaciones del Gobierno Vasco,

Pulgar Vernalte, F. & Marcos Maciá, S. (2008). Ondarenet: el archivo del patrimonio digital vasco. In X Jornadas de Gestión de la Información, Madrid (Spain), 20-21 November 2008

Ras, M & Van Bussel, S. (2007) Web archiving user survey. Retrieved april 2009 from http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf

Serra, E. (2006). Archivando la Web catalana: iniciativas cooperativas de preservación digital en Catalunya". In *La Recuperación de la memoria, muchas más oportunidades que realidades: el trabajo cooperativo de archivos, bibliotecas y museos. Universidad del País Vasco* 2006. Retrieved april 2009 from http://www.bnc.es/bc/archivando_web_catalana.pdf

UNESCO. (2003). Directrices para la preservación del patrimonio digital. Retrieved april 2009 from http://unesdoc.unesco.org/images/0013/001300/130071s.pdf