Patrimoni Digital de Catalunya, a year and a half experience

Ricard de la Vega Sivera, Natalia Torres, Joan Cambras

Centre de Supercomputació de Catalunya

Abstract. E-repositories are part of the e-science, and they are based on the e-infrastructure. The Centre de Supercomputació de Catalunya (CESCA) together with the Consorci de Biblioteques Universitàries de Catalunya started in 1999 a cooperative repository, named TDR, to file, in digital format, the full-text of the read thesis at the universities of our country in order to spread them worldwide in open access, while at the same time, preserving the intellectual copyright of the authors. Since then, four additional cooperative repositories have been created: RECERCAT for research papers; RACO for scientific, cultural and erudite Catalan magazines; MDC for Catalan digital collections of pictures, maps, posters and old magazines; and PADICAT for archiving Catalan digital web content;
The main objective of the latter is to archive Catalan web sites. That is, PADICAT collects, processes and provides permanent access to the entire cultural, scientific and general output of Catalonia in digital format. The repository manager is the Biblioteca de Catalunya, as the institution responsible for compiling, processing and distributing the bibliographic heritage of Catalonia, while CESCA is the technology partner.
On September 11th, 2006 the repository went into operation for the general public, with some thirty websites archived. After one year and a half, it has 2.720 captures of more than 1.000 websites. This includes 34 million files (HTML, images...) and two terabytes of data.
The objective of this paper is to present PADICAT and our experience developing and managing it. We describe the repository briefly, we explain the technology used to implement it and we comment our experiences during its first year and a half.

1 Introduction

The e-repositories, based on the e-infrastructure, are a transversal component of e-science. Researchers access journal articles, thesis, research papers, datasets, web content... for their research and after that, produce new science using, for example, grid e-resources. Finally, they share the new knowledge across the scientific community (paper, thesis, web content...). The e-repositories are the places to implement this cycle.

The Centre de Supercomputació de Catalunya (CESCA) together with the Consorci de Biblioteques Universitàries de Catalunya (CBUC) started in 1999 a cooperative repository, named TDR[1], to file in digital format the full-text of the read thesis at the universities of our country to spread them worldwide in open

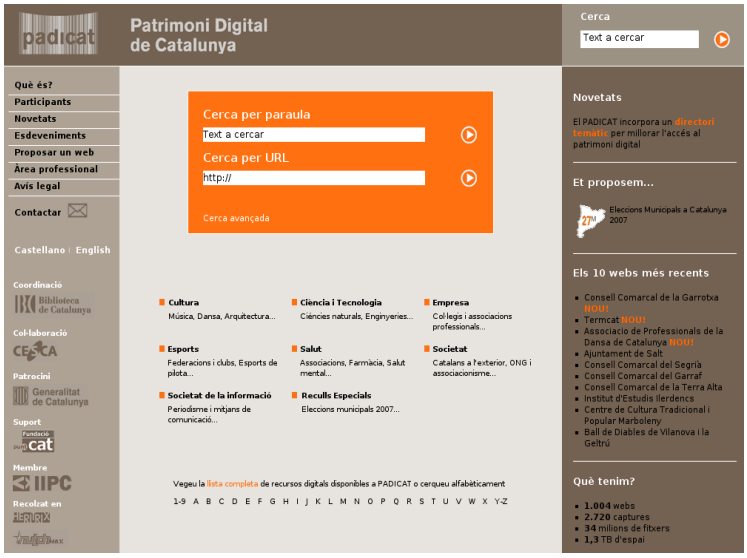

Fig. 1. PADICAT homepage

access preserving the intellectual copyright of the authors. Since then, four additional cooperative repositories[14] have been created: RECERCAT[2] for research papers; RACO[3] for scientific, cultural and erudite Catalan magazines; MDC[4] for Catalan digital collections of pictures, maps, posters and old magazines; and PADICAT[5] for archiving Catalan web sites.

The latter mentioned, Patrimoni Digital de Catalunya (PADICAT), collects, processes and provides permanent access to the entire cultural, scientific and general output of Catalonia in digital format.

The objective of this paper is to present PADICAT and our experience developing and managing it. First we describe the repository; second we explain the technology (hardware and software) used to implement it and finally, we comment on our experience during its first year and a half since it was put to operation to the general public.

2 The digital heritage of Catalonia

PADICAT[15] collects, processes and provides permanent access to the entire cultural, scientific and general output of Catalonia in digital format. In some countries, similar projects are known as national digital archives or web archives. The best known of these are the giant Internet Archive[6], Australian´s Pandora[7] and Sweden´s Kulturaw3[8]. PADICAT, as well as these other projects, is a member of the International Internet Preservation Consortium (IIPC) [9].

In accordance with the general trend among national libraries, the archive model used by BC is a hybrid system consisting of the following:

– Mass compilation of open-access digital resources published on the internet.
– Systematic archiving of the web site output of Catalan organizations.
– Fostering of lines of research through themed integration of the digital resources pertaining to specific events in Catalan public life.

The 21st of July 2006, work began on the automated collection of web sites that were candidates for becoming part of PADICAT. The first of these were those of the town councils of Berga and Palafrugell, and the professional associations of Quantity Surveyors and Technical Architects of Tarragona and of Social Workers of Catalonia. On September 11th, 2006 PADICAT went into operation for the general public, with some thirty web pages archived.

By 2009, PADICAT should be in an optimum position, whereby this system operates at full capacity, with quantitative indicators of 10,000 web pages captured in different editions. This may include some 50 million files and 30 terabytes of data.

3 Technology

A complex architecture (see figure 2) has been deployed for PADICAT: some nodes are exclusively dedicated to the project while others are shared with other repositories, all of them within a high availability cluster dedicated to e-information resources. Both have been installed specialized and additional software to meet the requirements of the repository.

3.1 Hardware

The service has four HP ProLiant DL360 G4p nodes responsible for the functions of collecting and indexing websites. These nodes are virtualized by Xen to better adapt resources according to need, for example, giving more memory to the collection when there is not indexing.

On the other hand, a Linux cluster of high availability with the required features load balancing and fault tolerance is responsible to search and display results on the web interface. This cluster is shared with other cooperative information e-repositories such as TDR, RECERCAT and RACO.

The nodes are connected via fibre to a Storage Area Network (SAN) and the system is complete with a robot which is stored on tape backups of the data.

3.2 Software

PADICAT is based on the application of a number of computer programs (see figure 3) that allow web pages published on the Internet to be collected, stored, organized, preserved and permanently accessed. Heritrix[10] is the Internet Archivess archival-quality web crawler that harvests and stores, in compressed files, the crawled web pages. Then, NutchWAX[11] generates the indexes that will then be used to search the data. Finally, there are two interfaces for searching and navigating the archived web document collections: the Web Archive Access (WERA)[12]

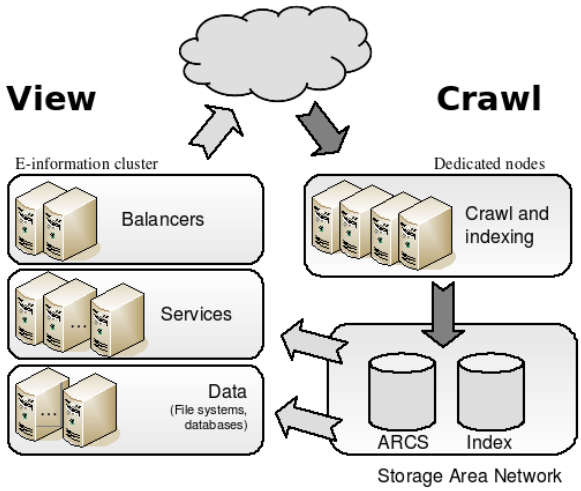

Fig. 2. The PADICAT architecture

is a freely available solution that allow keyword searches, and Wayback[13], that allow searches only for URL´s.

As partners in IIPC, sharing experiences with other centers also work for the preservation of digital heritage. Most projects are using the same software. While all are operational and used in production environments, they are constantly changing and lack of functionality in various aspects, such as documentation, planning and statistics.

All programs used in the project are free software[16], which allows its adaptation and the development of new modules that later can be contributing to the community. CESCA is working on a catalogue and statistics modules.

4 Experiences of the first year and a half

Currently, more than 1.000 webs are available for cooperative agreements with institutions representing catalan civil society, digital resources proposed by the public (blogs, entities, etc.), and webs related with special events made. It has 2.720 captures of these webs distributed in over 34 million files (see figure 4).

In this first year and a half, a wide selection of websites representing the catalan civil society, such as municipalities, universities, professional associations, cultural or sporting, political parties, companies and media, have already reached cooperation agreements with PADICAT, a total of about 290; also have received more than 350 proposals from other resources and have collected the digital resources associated with two events.

The space required to store the data has evolved in a linear manner to captures made, reaching now to 2 TB. Approximately 30% of the space is needed for queries

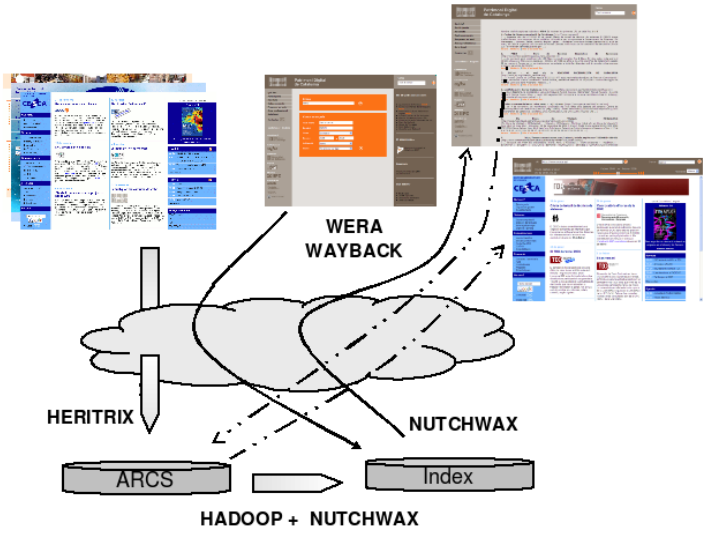

Fig. 3. The PADICAT software

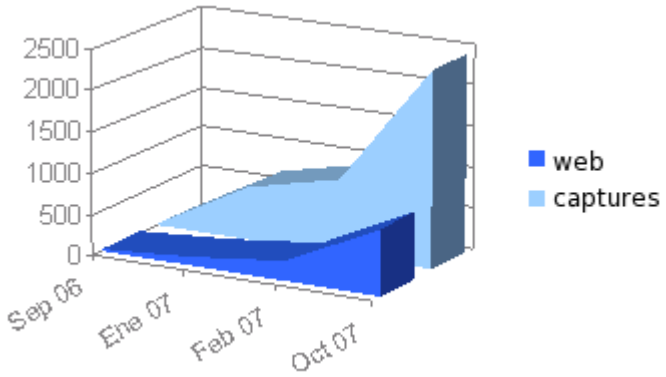

Fig. 4. PADICAT captures

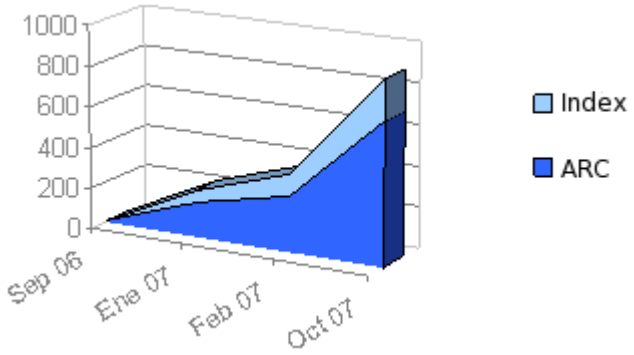

Fig. 5. PADICAT used space

with the collection, while the rest belongs to resources with ARC format (see figure 5). Besides this space, during the generation of indexes also temporary space is needed, specifically, more than double that finally took the indexs.

In early 2008 the repository has two new developments. On the one hand it has been incorporated Wayback, a search engine software for URLs which allows access to the specific capture typing the URL of the page. On the other hand, a directory that offers a thematic classification of web resources captured in seven areas of knowledge: culture, science and technology, sports, health, society, and information society, which are subdivided into 72 subcategories. The navigation structure is completed with an eighth category dedicated to special projects devoted to the election campaigns of 2006 and 2007. Thus the new directory, allows users to access with two "clicks" to the 62 resource classified.

4.1 Events

The events were intended to reflect the environment and the situation of catalan society regard to specific situations. Up to now have conducted three tracks:

Elections to Parlament de Catalunya The elections to Parliament were held in Catalonia on the 1st November of 2006. To understand this key event for the Catalan society, selected and harvested several versions of 76 online digital resources (political parties with and without parliamentary representation that presented candidature, personal blogs, candidates web pages, mass media, institutional digital resources, political foundations and research centers, etc) that have been incorporated to the Digital Heritage of Catalonia, to follow the evolution of the election campaign up in the Catalan Web.

Local elections in Catalonia 2007 Along the same lines as the special issue about the elections to the Catalan Parliament on the fist of November 2006, PAD-

ICAT has intensevely monitored the electoral campaign for the local elections of 27st May. Information was archived from the blogs and webs of principal candidates for the mayor's office of 48 Catalan municipalities (i.e. regional capitals, towns with more than 100,000 inhabitants and towns with more inhabitants than the regional capital itself).

Altogether these include the webs and blogs of 278 candidates, more than 100 local digital media and 30 media covering Catalan issues, etc. The mayoralty candidates for the municipalities mentioned above from parties that are represented in the Parliament, and from parties, local or not, that already had local representation were monitored. In this last case, monitoring of candidates was limited to municipalities where the parties already had representation and was not extended to others where the candidates could run for election since the party was not represented prior to the elections. Finally, we have also monitored those parties that, after the elections, have obtained at least one councillor for the first time.

The aim is to give a global view of the campaign's development on the Catalan web, wich could and should enable future studies in this key area of Catalan society to be carried out.

Folk-rock Internet has become a new mean of music distribution where the bands broadcast their sonorous proposals. The proliferation of the music in the net involves, some times, the absence of this same music in a physical format as, for example, a CD, and therefore, it has not been edited by any Record company. Without this traditional editing phase, regulated by the law of Dipsit legal (Legal Deposit), an important part of the musical richness of our country could disappear, attending to the permanent process of change or even renovation of contents in any web page.

It could be a difficult process, sometimes, to assign a label to that style of music you are listening in a concret moment. And folk-rock is not an exception to that. For us, this label envolves that kind of music that uses the sonorities of the traditional catalan music mixed with modern music sounds, that is, bands that uses a gralla or a tenora together with an electric guitar and drums.

This next selection of digital resources has been elaborated by Jordi Soler Sala, thanks to the collaboration between the Biblioteca de Catalunya (National Library of Catalonia) and the Escola Superior de Msica de Catalunya (ESMUC). Contains different examples of bands from the catalan speaking and culture territories, which uses Internet as a mean to diffuse their music, and other digital resources of all kinds, as music festivals, specialized media, record companies or management companies, where this kind of music is present.

In addition to these, alse it´s preparing other that will be available during the first semester of 2008. It will follow the guidelines of the first two about elections, but in this case, for the Spanish general election of 9th of march of 2008.

5 Conclusion

On September 11th, 2006 PADICAT was open to the general public, with some thirty websites archived. One year and a half after it has 2.720 captures of more than 1.000 web sites including 34 million files (HTML, images...) and two terabytes od data.

This first year and a half, from the technological point of view, the number of nodes have been doubled and the lack of maturity of some of the software have been seen. To overcome this latter aspect we are working on the development of new features that may contribute to the free software community. Currently, a catalogue module and other of statistics is been built.

By 2009, PADICAT will be in an optimum position, whereby this system —a pioneer in Spain and a benchmark in Europe— operates at full capacity, with quantitative indicators of 10,000 websites archived in different editions. This may include some 50 million files and 30 terabytes of data. Furthermore, cooperation agreements are scheduled to be signed with 500 institutions of all kinds and online open access to a considerable part of the collection will be available.

Acknowledgements

The developing of PADICAT would not have been made possible without the collaboration of many people, from BC and from our consortia. Thanks to the BC staff: Eugènia Serra, Ciro Llueca, Núria Rubio, Daniel Cócera and the positioning metadata staff. From CESCA, thanks to Jordi Prats and Xavier Torelló. Externally, thanks to all the people involved in the project, both the institutions with cooperation agreements and the people that has recomend resources to preserve.

References

1. Tesis Doctorales en Red (TDR), *http://www.tesisenred.net*.
2. Dipòsit de la Recerca de Catalunya (RECERCAT), *http://www.recercat.cat*.
3. Revistes Catalanes amb Accés Obert (RACO), *http://www.raco.cat*.
4. Memòria Digital de Catalunya (MDC), *http://www.cbuc.cat/mdc*.
5. Patrimoni Digital de Catalunya (PADICAT), *http://www.padi.cat*.
6. Internet Archive (IA), *http://www.archive.org*.
7. Australia's web archive (PANDORA), *http://pandora.nla.gov.au*.
8. Kulturalw3, *http://www.kb.se/kw3/ENG*.
9. International Internet Preservation Consortium (IIPC), *http://netpreserve.org*.
10. Heritrix, *http://crawler.archive.org*.
11. NutchWAX, *http://archive-access.sourceforge.net/projects/nutch*.
12. WERA, *http://archive-access.sourceforge.net/projects/wera*.
13. Wayback, *http://www.archive.org/web/web.php*.
14. Huguet, Miquel; Anglada, Lluís; Vega Sivera, Ricard de la. 1st Iberian Grid Infrastructure Conference Proceedings (IBERGRID). Santiago de Compostela. Spain. May 14-16, 2007. Santiago de Compostela: Centro de Supercomputacin de Galcia, 2007. "Catalan Policies and Experiences on Cooperative Repositories", p. 63-75.

15. Llueca, Ciro. Archivando la Web, el proyecto PADICAT (Patrimonio Digital de Catalunya), *El profesional de la informacin*, **15**, n. 6, pp. 473-478, (2006).
16. Vega Sivera, Ricard de la. Software libre en repositorios de e-informacin, *El profesional de la informacin*, **17**, n. 1, pp. 49-55, (2008).