Patrimoni Digital de Catalunya, a year and a half experience

R. De la Vega, N. Torres, J. Cambras

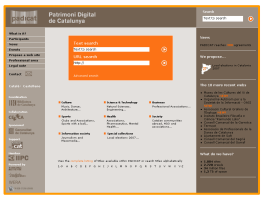
Centre de Supercomputació de Catalunya

Introduction

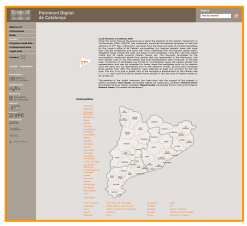
PADICAT collects, processes and provides permanent access to the entire cultural, scientific and general output of Catalonia in digital format. The repository manager is the Biblioteca de Catalunya, as the institution responsible for compiling, processing and distributing the bibliographic heritage of Catalonia, while CESCA is the technology partner.

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.



PADICAT offers a thematic classification of web resources captured in seven areas of Knowledge that allows users to access with two "clicks" to resources.

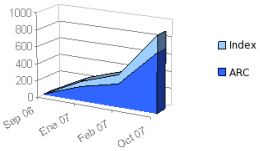


Page dedicated to special projects devoted to the election campaigns 2007.

What do we have?

- 1,004 sites
- 2,720 crawls
- 34 million files
- 1.3 TB of space
- 300 agreements
- 3 collections

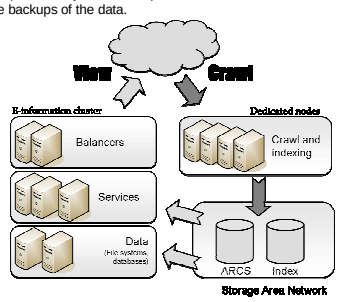
Used space



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 with the collection, while the rest belongs to resources with ARC format.

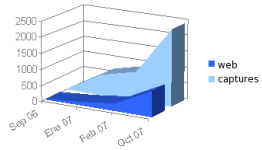
Hardware

- 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.
- 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.
- 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.



Some resources crawled available from PADICAT website.

Crawled resoures



PADICAT has available 2,720 captures of more than 1,000 webs.

Experiencies

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. 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 three events.

The events were intended to reflect the environment and the situation of Catalan society regarding 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

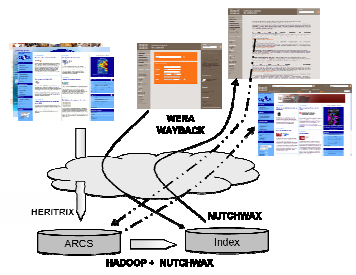
Along the same lines as the special issue about the elections to the Catalan Parliament on the fist of November 2006, PADICAT has intensevely monitored the electoral campaign for the local elections of 27st May. Information was archived from the blogs and webs of principal candidates for the mayor's office of 48 Catalan municipalities.

- Folk-rock

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. 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.

Software

- HERITRIX: an archival-quality web crawler that harvests and stores, in compressed files, the crawled web pages.
- NUTCHWAX: generates the indexes that will then be used to search the data.
- WERA: interface that allows keyword searches and navigating the archived web document collections.
- WAYBACK: interface for searching that allow URL searches and navigating the archived web document collections.



Conclusion

On September 11th, 2006 PADICAT was open to the general public, with some thirty websites archived. Eighteen months after, it has 2,720 captures of more than 1,000 web sites including 34 million files (HTML, images...) and two terabytes of 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.

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.

More information:

padicat
PATRIMONI DIGITAL DE CATALUNYA

http://www.padi.cat

Member of IIPC