CAT (CURATOR ARCHIVING TOOL): IMPROVING ACCESS TO WEB ARCHIVES

Ciro Llueca, Daniel Cócera
Biblioteca de Catalunya (BC)
Barcelona, Spain
padicat@bnc.cat

Natalia Torres, Gerard Suades, Ricard de la Vega
Centre de Supercomputació de Catalunya (CESCA)
Barcelona, Spain
padicat@cesca.cat

ABSTRACT

PADICAT is the web archive created in 2005 in Catalonia (Spain) by the Biblioteca de Catalunya (BC), the National Library of Catalonia, with the aim of collecting, processing and providing permanent access to the digital heritage of Catalonia. Its harvesting strategy is based on the hybrid model (massive harvesting of .CAT top level domain; selective compilation of the web site output of Catalan organizations; focused harvesting of public events). The system provides open access to the whole collection, on the Internet. We consider necessary to complement the current search and visualization software with a new open source software tool, CAT (Curator Archiving Tool), composed by three modules aimed to managing effectively the processes of human cataloguing; to publish the digital resources on directories and special collections; and to offer statistical information of added value to end users. Within the framework of the International Internet Preservation Consortium meeting (Vienna 2010), the progresses in the development of this new tool, and the philosophy that has motivated his design, are presented to the international community.

1. INTRODUCTION

PADICAT (in english Digital Heritage of Catalonia) is the web archive created in 2005 in Catalonia (Spain) by the Biblioteca de Catalunya (BC), the National Library of Catalonia, and the Centre de Supercomputació de Catalunya (CESCA), with the aim of collecting, processing and providing permanent access to the digital heritage of Catalonia, understood as the entire cultural, scientific and general output of Catalonia in digital format[1] and published on the Internet. The goal of PADICAT[2] is to archive the Catalan Internet.

The BC is member of the IIPC (International Internet Preservation Consortium) and, as in the rest of ongoing projects, the web archive placed in Barcelona is based on the application of several IT software that allows the crawl, storage, preservation and the permanent access to a series of versions of web pages published by the Catalan community[3], in Spain.

From the initial analysis of the Internet Archive, Kulturarw3 (National Library of Sweden), Pandora (National Library of Australia), and Netarkivet.dk (The Royal Library, State & University Library, Denmark) models, and according to what we consider the common trend among national libraries and archives, the model of PADICAT's repository is the hybrid one, consisting of:

- Mass compilation of open-access digital resources published on the Internet, through the exhaustive harvesting of .CAT top level domain.
- To stimulate the systematic archiving of the web site output of Catalan organizations, through its identification and the signing of cooperation agreements with the entities and companies representative of the Catalan society.
- Fostering lines of research through themed integration of the digital resources pertaining to specific events in Catalan public life, such as electoral campaigns on the Internet, Catalan music on-line or museums of Catalonia.

PADICAT uses in his daily functioning Heritrix, Nutchwax, Wera, Wayback Machine, as well as Web Curator Tool software (WCT), developed by the National Library of New Zealand in collaboration with the British Library.

[1] Webb, C. *Guidelines for the Preservation of Digital Heritage*. United Nations Educational, Scientific and Cultural Organization, Paris, 2003. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

[2] See What PADICAT is at http://www.padicat.cat/en/quees.php

[3] Gomes, D.; Silva, M. J. "Characterizing a National Community Web". *ACM Transactions on Internet Technology*, vol 5, num 3 (Aug 2005). http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf

In September, 2010, PADICAT offers through his web site (http://www.padicat.cat/en/) open access to 53.249 crawls from 30.481 web sites and keep cooperation agreements with 442 public administrations, companies, universities, professional associations and cultural centers of Catalonia. The open access offered to the whole collection, through searching or browsing, is one of the strong points of the Catalan web archive. Together with the goal of preserving the published information on the Internet, visualization of the crawled collection through Internet has become a priority: being aware of the legal restrictions that concerns the majority of existing web archives, PADICAT has followed, from the beginning, the premise of the open access to his collection, premise defended by the Internet Archive, at the same time that has been weaving manually a network of entities and companies that allows, thanks to the signing of cooperation agreements, to disseminate the crawled versions through web archive's site.

The strategic goal of guarantee an optimum retrieval and visualization of the crawled information has been carried out through the assignment of added value to the crawled resources data, associating new descriptive information to the digital articles. In short, cataloguing the web resources that PADICAT has been harvesting and those that has to harvest soon. At the present time, 17.700 sites have been catalogued.

The information provided by manual cataloguing will have to allow an essential improvement in the identification of the resources and its retrieval according to a standardized title, the alphabetical positioning in the thematic directory of the repository, its belonging and indexing inside thematic concrete categories, the interrelationship between the resources inside the PADICAT collection, etc. The following challenge will be the modification of the scoring of the current search systems (Wera and Wayback) considering the metadata coming from the cataloguing.

Definitively, new search capabilities that will have to offer as a final result a better visibility of those resources that make up the collection, allowing users to interrogate the system without the need of knowing the URL of the desired resource. The cataloguing of web reources has stimulated the design and the production of a set of tools that we call CAT (Curator Archiving Tool).

The aim of this report distributed at the International Internet Preservation Consortium (sited in the 7th International Conference on Preservation of Digital Objects, iPRES2010) is to show to the international community the improvements raised by PADICAT staff on the processes of retrieval and visualization of the information contained on web archives, as well as to put at the public disposal for the equivalent projects the advances in these lines of research and implementation.

2. CAT (CURATOR ARCHIVING TOOL)

It seems evident that harvesting process is well covered by quite mature software, such as Heritrix. However, indexing software and the corresponding modules for visualization have not evolve at the same pace.

As in the rest of catalogues, the retrieval of information contained in web archives is carried out through searching or browsing, understood as navigation using a thematic directory. Only the searching part has specific software quite drawn to the community, such as WERA and Wayback, searching by URL or keyword query. Navigation through thematic directories, as used in web archives like Pandora or UK WAC, requires the traditional process of cataloguing, like it's being carried out also in PADICAT, with the final goal of integrating his collections to the BC's catalogue.

Therefore, the steps of the process of web site resources managements, such as the harvesting or searching, have mature software and are vastly used in similar projects, but there is a gap in open source software for the cataloguing and publication of these resources.

In front of this challenge, and to guarantee a better retrieval and visualization by non-traditional web archive users, PADICAT staff have developed several implementations and carried out various tests since 2005, that have made possible, nowadays, to launch a tool addressed to influence positively the processes of searching and information retrieval, as well as in the presentation of this information to users consulting a web archive.

These are the modules that make up CAT (Curator Archiving Tool), thought taking into account criteria of modularity, scalability and internationalization, in order to their publication as open source software.

It has been designed software made up by three modules: the basic one, based on the cataloguing process (MOCA), complemented with more functionalities from the additional modules: publication (MOPU), and statistics (MOST), all three of them make up the software CAT (Curator Archiving Tool), object of the present communication.

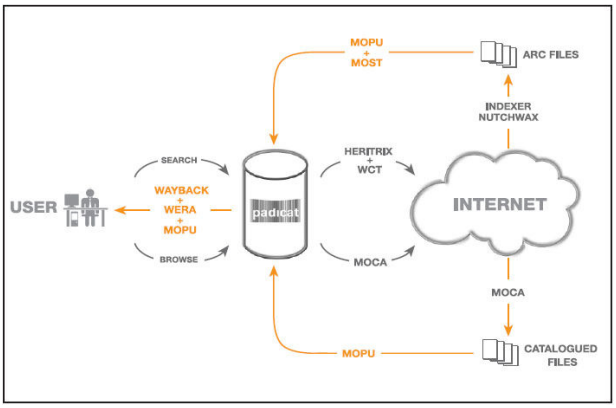

Figure 1. Software used at PADICAT

A description of the modules that make up CAT (Curator Archiving Tool) is explained next.

2.1. MOCA (Cataloguing Module)

The cataloguing of resources started using already existing tools, essentially pioneering Web Curator Tool (WCT). With the experience lacks were detected in the current software. Also some particular needs from the project are not covered neither. To solve it, a development of a new tool focused only on the cataloguing process was considered.

One of the main MOCA's requirements is that it should allow to record a changelog in the metadata of the resources catalogued because this information might be relevant: a change in the associated URL of the preserved resource is a clear case to take into account.

The information collected to catalog the resources is based on Dublin Core metadata model. It can be broadened easily without altering the data model. In PADICAT crawled resources are processed following an indexing system[4] of Biblioteca de Catalunya, which is based on a thematic classification specific to the PADICAT web portal, and also according to the origin of the guideline crawl (because it belongs to .cat domain, collaboration agreement websites, proposed resources or monographic collections).

An authentication system allows only to log in on the application the users identified with user and password, and once validated the user only can access to the actions that he can carry out according to the assigned roles. The use of roles simplifies the user pages and allows controlling accurately the actions that each of them can carry out. Three roles have been defined based on the professional profiles of the staff involved in the cataloguing process. These roles are manager (maintenance tasks of the tool), cataloguer (inserts and updates the catalogued resources) and observer (control of the catalogued resources).

It is mandatory to check in MOCA's workflow if a resource already exists in the catalogue in order to avoid duplicated resources. Once passed this first step then the minimum metadata required has to be introduced and then saved.

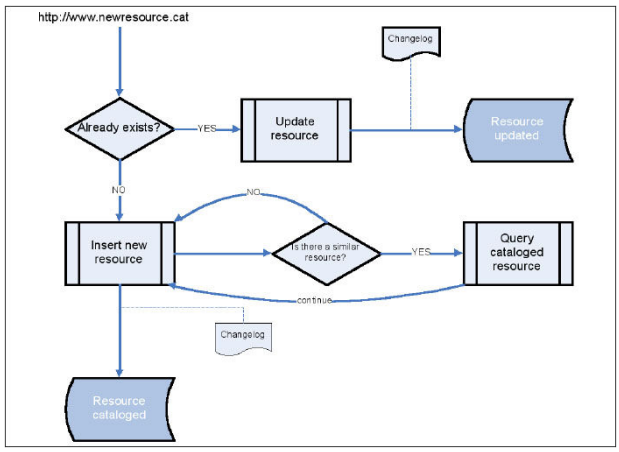

Figure 2. MOCA workflow

Some screens of the main interfaces are shown below: the interface allows introducing the metadata associated to a resource, searching interface allows making combined searches using any field of metadata and batch update interface allows updating multiple resources at the same time.

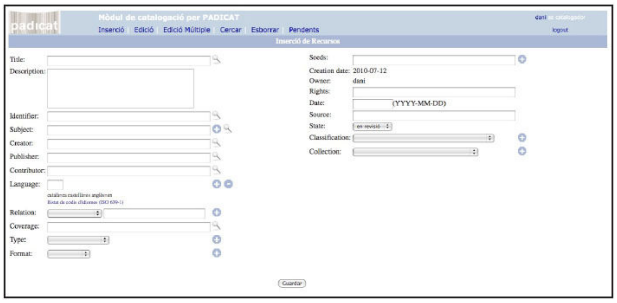

Figure 3. MOCA insert resources screen shot v1.0

[4] LEMAC, *Llista d'encapçalaments de matèria en català*, adaptation of the LCSH, *Library of Congress Subject Headings* with contributions proceeding from Laval RVM, *Répertoire des Vedettes-Matière*, and RAMEAU, *Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié*, that BC maintains.

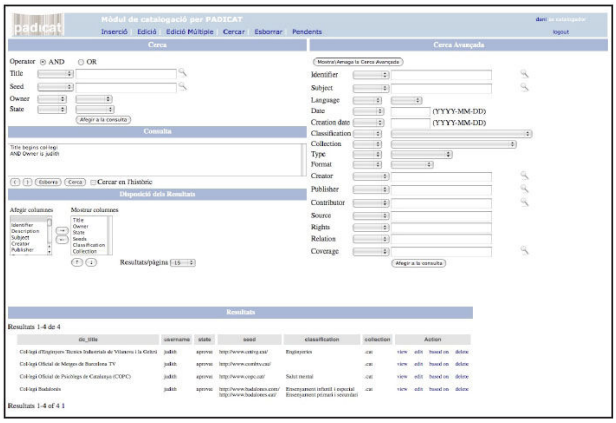

Figure 4. MOCA search resources screen shot v1.0

Once this tool has been developed several scripts were created to allow the migration of the existent information in WCT to the new MOCA's data model. The release date of MOCA's 1.0 version is scheduled to the end of March 2011.

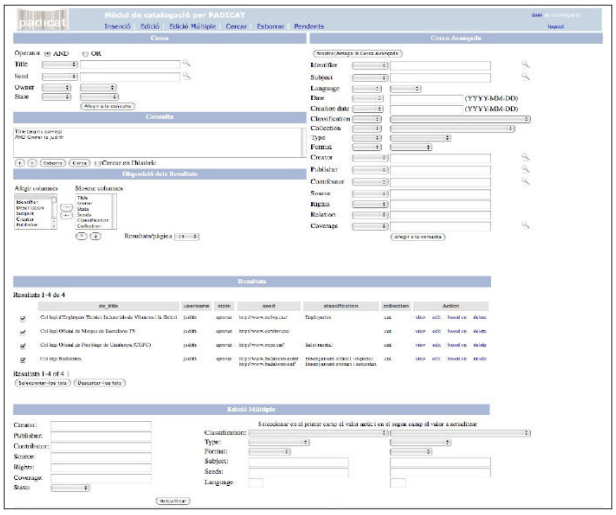

Figure 5. MOCA batch update screen shot v1.0

2.2. MOPU (Publication Module)

MOPU will automatically generate the thematic directory where all resources that were previously catalogued using MOCA and also were selected to be published will be available.

Optionally each resource will have an attached item information page with descriptive information arising from the cataloguing metadata that were considered relevant for the visualization of the crawled resource. Metadata included in the item information page are selected in individualized and customized method for each resource. In order to use MOPU, it is mandatory to have MOCA properly installed.

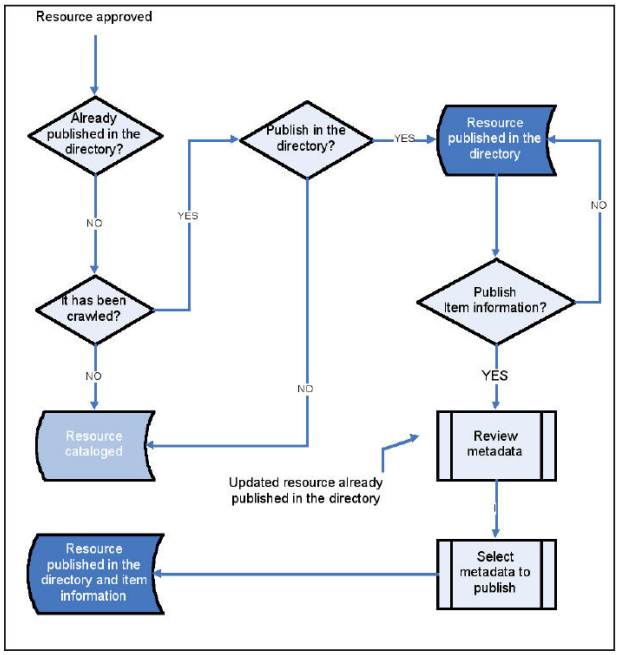

Figure 6. MOPU workflow

MOPU needs the definition of a new role, known as validator. This new user role will be in charge of the management of the whole publication process. Validator will check the metadata inserted by cataloguers and he will also give his approval to the publication of the resource in the thematic directory. Moreover, if necessary he will select which metadata will be shown on the item information page. If the validator detects any deficiency on metadata, he will mark the resource so as to be checked by cataloguers. Additionally, through an attached note field to the resource the validator will write down the reason why the metadata needs to be changed.

In the same way as MOCA, in MOPU a changelog that will allow tracking the publication process and the changes in the item information page will be automatically generated.

Some screens of the main interfaces of the prototype pilot are shown below: management interface of the thematic directory according to categories and metadata selection interface and the interface for the selection of metadata to include in the item information page.

Figure 7. MOPU prototype pilot screen shot

Figure 8. MOPU prototype pilot screen shot

2.3. MOST (Statistics Module)

This module will allow generating statistics on the information included in the repository from the reports generated by the crawling software, such as number and type of files belonging to each preserved crawl, size and frequency of each crawl, etc. This is an added value service because PADICAT offers searching and browsing through all crawled resources in open access.

Binding this module with MOCA and MOPU will allow including statistics on the resource item information page.

2.4. Roadmap

As soon as a mature enough and stable version of the CAT (Curator Archiving Tool) modules is reached, with feedback on their use, it will be released to the community as open source software. Release date for MOCA is expected to be before the end of March 2011.

3. CONCLUSIONS

Harvesting web sites published on the Internet is covered enough by quite mature software, like Heritrix. However, indexing software and the corresponding modules for visualization have not evolved at the same pace. For web archives like PADICAT, that offers open access on the Internet to his collection, is essential to improve the processes of search and visualization.

The conception of the software CAT (Curator Archiving Tool), allows improving the search and visualization of the resources preserved in web archives, thanks to his contributions in:

- Human impact in describing and indexing through the Cataloguing module.
- Automated generation of directories and special collections through the Publication module.
- Increase of the offer of statistical data through the Statistic module.

For the web archive managers, the interaction of the three modules will allow saving time thanks to the automation of processes that are manual at the moment:

- Automated integration of preserved resources in alphabetical lists, in thematic directories or in special collections that make up the access to the collection through navigation.
- Batch update of catalogued resources sharing the same characteristics.
- Improvement of workflows between different roles that operate in cataloguing and publication of resources at web archive's site.
- To follow the strategy of Biblioteca de Catalunya (BC), with the goal of the integration of the preserved resources into catalogues and ordinary search systems of the Library.

For the web archive users, the creation and starting of the software CAT (Curator Archiving Tool) will allow an improvement on the search capabilities and the visualization of preserved resources:

- Search by thematic navigation, more effectively and exhaustively, complementing the current systems of searching by keyword or URL.
- Search by keyword in a more appropriate way, once modified the current search system scoring with the metadata proceeding from cataloguing.

- Integrated search, once the PADICAT collection has been integrated into the catalogues and ordinary search systems of the Biblioteca de Catalunya.
- Access to added value information for each preserved resource, coming from human cataloguing or from system-generated statistics.

For the international community, the publication as open source software of the three modules of CAT (Curator Archiving Tool) allows to:

- Contribute to the goal of acquire, preserve and make accessible the Internet information for future generations, around the world, promoting the global exchange and international relations.
- Foster human intervention in cataloguing and publication processes in web archives to improve the access and visualization of these systems.
- Achieve public and political appreciation to the usefulness of Internet digital preservation projects, as the web archives presents at the IIPC annual meeting are.