The National Library of Catalonia and the Digital Heritage

Ciro Llueca
cllueca@bnc.cat

Biblioteca de Catalunya
www.bnc.cat

Digital Heritage
Unesco 2003

Digital heritage embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other information **created digitally**, or **converted into digital** form from existing analogue resources.
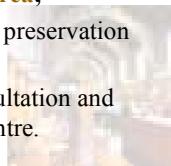
.

National Library of Catalonia
Strategy

The mission of the Biblioteca de Catalunya is to

collect, to preserve and to diffuse

the Catalan bibliographic production and that related to Catalan linguistic area;

to watch over the bibliographic heritage preservation and divulging;

and to keep up the condition of a consultation and universal scientific research centre.

Bibliographic production at 2007

Books, Leaflets, Printed sheets, Periodic publications, Engravings, Maps and plans, Posters, Postcards, "Slides", Prints or sound recordings, Movies...

+

Websites: e-books, e-News, Weblogs, Webcams, on-line votes, Chats, Corporative websites, e-commerce, Personal webs, Digital articles and papers, software documentation...

National Library of Catalonia
Projects focused in digital heritage

Retrospective
- ARCA (Old Serials Repository)
- * RACO (Open Serials Repository)
- CLACA (Catalan Classics Portal)
- Google Book Search

Born digital
- PADICAT (Catalan Web Archive)

http://mdc.cbuc.cat/

National Library of Catalonia
ARCA (Old Serials Repository)

ARCA (Arxiu de Revistes Catalanes Antigues)
- Full online access & search
- Digitalizing from closed collections
- 50 titles, 50.000 pages (goal 2009: 300 tit.)
- Cooperating with several public & univ libraries
- Copyright: Public Domain

http://www.bnc.cat/digital/arca/

arca

National Library of Catalonia
RACO (Open Serials Repository)

* RACO (Revistes Catalanes amb Accés Obert)

Full online access & search

Digitalizing and Publishing in pdf the "press copy" from live collections

100 titles

Cooperating with university libraries

Copyright: Agreements with publishers (universities, professional associations, etc.)

http://www.raco.cat/

National Library of Catalonia
CLACA (Catalan Classics Portal)

CLACA (Clàssics Catalans)

Full online access & search

Digitalizing from own holdings (books, articles, letters, graphics, audio & video, etc.)

Thematic collections: Jacint Verdaguer (poet)

Copyright: Public Domain

http://www.bnc.es/fons/claca.php

National Library of Catalonia
Google Book Search

Google Book Search

Full online access & search

X00.000 titles

Cooperating with 4 Libraries in Catalonia

Copyright: Public domain

http://books.google.com/

National Library of Catalonia
PADICAT (Catalan Web Archive)

PADICAT (Patrimoni Digital de Catalunya
= Digital Heritage of Catalonia)

Full online access & search

100.000 websites (goal 2009)

Cooperating with a technical partner (CESCA), and 1.000 web producers (goal 2009)

Copyright…

http://www.padicat.cat/

Born digital heritage is ephemeral

1999   2001   2003   2006

Experiences in Digital preservation

Germany, 1997	Iceland, 1997
Australia, 1996	Japan, 2002
Austria, 1999	Lithuania, 2002
Canada, 1994	Norway, 2001
Denmark, 1998	New Zeland, 1999
USA, 2000	Netherlands, 1995
Estonia, 2004	Quebec, 2000
Finland, 1997	United Kingdom, 2004
France, 2000	Czech Republic, 2001
Greece, 2003	Sweden, 1996

netpreserve.org

Experiences in Digital preservation

**1996 →
55 billion Web pages**

Internet Archive http://www.archive.org/

Experiences in Digital preservation

**1996 →
Selective archive
26.000 webs, 33 million files**
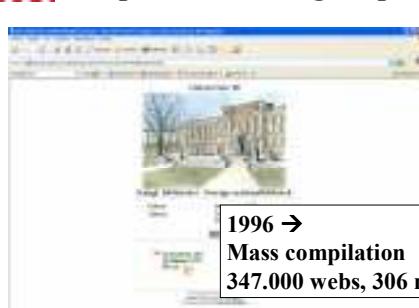
Australia http://pandora.nla.gov.au/

Experiences in Digital preservation

**1996 →
Mass compilation
347.000 webs, 306 million files**
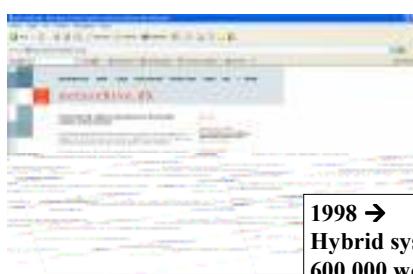
Sweden http://www.kb.se/kw3/ENG/

Experiences in Digital preservation

**1998 →
Hybrid system
600.000 webs aprox.**

Denmark http://netarchive.dk/index-en.php

Models: advantages & inconveniences

Exhaustive or integral
Rich and appropriate collection
Short-time evident results
Low cost
Excludes invisible Internet
Irregular collection
Limited access to results

Selective
Well-balanced collection
Maximum access to results
Strategic model
Partial view
Out of context
High cost

Hybrid
Rich and well-balanced collection
Appropriate
Strategic model with evident results
Includes invisible Internet
Maximum access (in regard to author's copyright)
High cost, similar to Selective model

PADICAT
Strategy

The mission of PADICAT project is

to collect, to preserve and to diffuse

the catalan **digital** production and that related to catalan linguistic field;

to watch over the digital heritage preservation and divulging;

and to keep up the condition of a consultation and universal scientific research centre.

PADICAT
Exhaustive captures

- Webs under .CAT domain
- Webs located in Catalan hosts
- Webs in Catalan language under other domains
  (.ES, .ORG, .NET, .COM, .INF, ...)
- Webs created by Catalan people
- **Other web sites** related to Catalonia, not included by the previous criteria
- **472 webs * 917 captures * 9,5 M files * 235 GB**

PADICAT
Selective captures

- Selection of 1.000 representative institutions of the Catalan society
  - City councils and local governments
  - Political Parties and Trade Unions
  - Professional Associations
  - Universities and Culture
  - Business

170 agreements from 09/11/2006

Ajuntament d'Arenys de Mar
Associació de Mestres Rosa Sensat
Universitat Politècnica de Catalunya
Col·legi de Pedagogs de Catalunya
Federació Catalana de Cineclubs
Ajuntament de Cerdanyola del Vallès
Ajuntament de Cervera
Col·legi de Biòlegs de Catalunya
Ajuntament de Girona
Fundació Joan Brossa
MACBA (Museu d'Art Contemporani de Catalunya)
Ajuntament de Manlleu
Tricicle Companyia Teatral S.L.
Casal Lambda
Ajuntament de Mollet del Vallès
Reial Club Deportiu Espanyol
Ajuntament de Montcada i Reixac
Col·legi de Periodistes de Catalunya
Ajuntament d'Olot
Universitat de Barcelona
Dagoll Dagom
Ajuntament de Ripoll
Ajuntament de Sabadell
Col·legi de Fisioterapeutes de Catalunya
Universitat Pompeu Fabra
Vilaweb
Ajuntament de Tarragona
Museu Episcopal de Vic
Ajuntament de Terrassa
Ajuntament de Torroella de Montgrí
Universitat Oberta de Catalunya
Ajuntament de Tremp
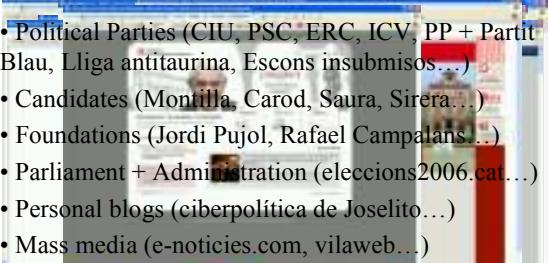Ajuntament de Vilanova i la Geltrú
Associacions
Casal Lambda

PADICAT
Focalized Action: elections 2006

- Political Parties (CIU, PSC, ERC, ICV, PP + Partit Blau, Lliga antitaurina, Escons insubmisos…)
- Candidates (Montilla, Carod, Saura, Sirera…)
- Foundations (Jordi Pujol, Rafael Campalans…)
- Parliament + Administration (eleccions2006.cat…)
- Personal blogs (ciberpolítica de Joselito…)
- Mass media (e-noticies.com, vilaweb…)

652 captures from 83 different webs

PADICAT
Repository and description

- Description, through a permanent identifier and Dublin Core metatags (descriptives and techniques)
- Repository, in ARC format
- Preservation, through the current and future technologies (migration, emulation…)

96% of files belong to
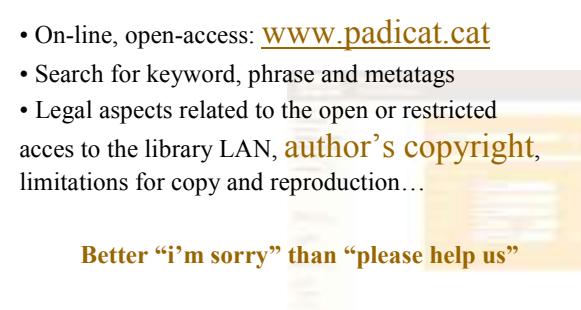
standard formats (html, jpeg, gif, pdf…)

PADICAT
Access

- On-line, open-access: **www.padicat.cat**
- Search for keyword, phrase and metatags
- Legal aspects related to the open or restricted acces to the library LAN, **author's copyright**, limitations for copy and reproduction…

Better "i'm sorry" than "please help us"

PADICAT
Highlights

- Beginning: June 2005, project: 2006-2008
- Hardware: 4 servers ProLiant DL360 G4p, Robot Scalar i2000, 10 TB (+ annual growth 10 TB)
- Software: Heritrix, BAT, NutchWax, Wera, Wayback
- Budget: 1.000.000 € (2006-08)
- Members of Int. Internet Preservation Consortium
- Professionals involved: 7 (librarians/technicians) + BC and CESCA teams,

  the Centre de Supercomputació de Catalunya

PADICAT
January 2009

- Pioneer project in Spain and reference in Europe (*regional* web) and Latin-America
- Several editions from 100.000 websites
- 50 million files, in 30 TB
- Agreements with 1.000 institutions
- Online acces to the major part of the collection

PADICAT
April 2007

Thanks for your attention!
www.padicat.cat

Biblioteca de Catalunya
Ciro Llueca, cllueca@bnc.cat