Design and selection criteria for a national web archive

Daniel Gomes
Sérgio Freitas
Mário J. Silva
University of Lisbon

The digital era has begun

- The web is the biggest source of information ever built
  - All kinds of publications: news, technical forums, books
  - There is information exclusively online
- Web data is ephemeral
  - Future generations will witness an information gap
- Need for web archiving
  - Historical purposes
  - Valuable data sets for research

Requirements of web archiving

- Conventional archiving requires strong human intervention
  - Cannot cope with the dimension of the web
- Automatic collection and storage
  - Minimal human intervention
- Expensive at large scale
  - Internet Archive

National web archives

- Divide to conquer
  - Each country archives its own web
- Need for selection criteria to define national webs
- Need for system architectures and specific software to support web archiving

Outline

  - Introduction
- Selection criteria to populate a national web archive
- The Tomba web archive prototype
- Conclusions

Methods to populate a web archive

- Delivery: publishers send contents to the web archive
  - Expensive for publishers
  - Hard to impose
  - Lack of standards and tools
- Harvesting: web archive gathers contents from the sites of the publishers
  - More load on the web archive
  - Hard to define selection criteria

Selection criteria for a national Search Engine

- Objective: provide relevant and up-to-date search results
- Broad selection criteria
  - Contents under .PT
  - Contents linked from .PT and written in Portuguese
- Relies on ranking mechanisms to exclude irrelevant contents from search results
- Web collection is periodically refreshed

Selection criteria for a national Web Archive

- Objective: preserve web data for historical purposes
- Narrower selection criteria
- Web collection is built incrementally
  - Save on storage
  - Preservation requirements
- Which selection criteria should be adopted?

Evaluating selection criteria for a national web archive

- Baseline: crawl for the tumba! search engine
- Selection criteria derived from web archiving requirements
  - More restrictive than s.e. selection criteria to avoid wasting resources on archiving irrelevant contents.
    1. ccTLD
    2. Media types
    3. Blogs
    4. Robots Exclusion Protocol

1. Restrict to ccTLD

- Easy to establish
- Country code Top Level Domains have a national scope
  - www.tumba.pt, .PT is the ccTLD for Portugal
- People also use gTLD (.com, .net, .org): commercial reasons, cheaper, faster to register.
- 49% of the Portuguese web is under .PT

2. Select media types

- Publication formats change with time but must be contents must be preserved
  - TXT->HTML->XHTML->?
- Preservation strategies according to media type
  - Conversion for open formats
  - Emulation for proprietary formats
- Preservation costs increase with media type diversity

Media type distribution

MIME	Avg. Size (KB)	%contents
text/html	24	61.2%
image/jpeg	32	22.6%
image/gif	9	11.4%
application/pdf	327	1.6%
text/plain	102	0.7%
Others	-	2.5%

Preserving HTML, JPEG and GIF covers 95% of the Portuguese web

3. Should blogs be archived?

- Historical relevance
  - Teenager who communicate with friends
  - Blog of the next President
- Identified by “blog” on the site name
- 15.5% are blogs
  - 63% under blogspot.com
  - 33% hosted under .PT

4. Ignore exclusion mechanisms

- Web archives use robots to gather contents
- Publishers may forbid harvesting (Robots Exclusion Protocol)
- Public interest overcomes private interest
- Robots exclusion protocol
  - 19.5% of the sites contained the robots.txt
  - 0.3% forbade the crawl
  - Might avoid crawlers from getting trapped

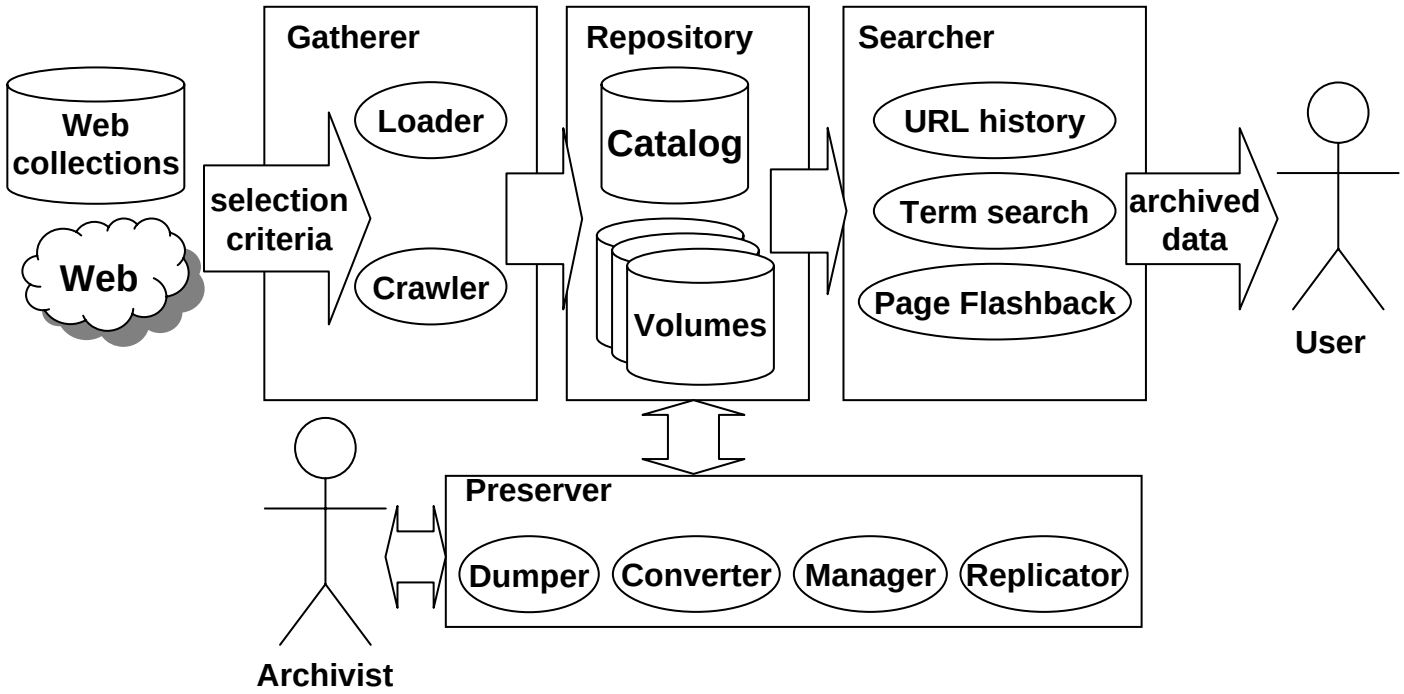
Outline

  - Introduction
  - Selection criteria to populate a national web archive
- The Tomba web archive prototype
- Conclusions

Requirements of Tomba

- Meta-data to enable interpretation and preservation
- Collection of contents built through incremental crawls
- Duplicates management
- Accessible by humans and machines
- Tools to manage and preserve

Tomba architecture



Tomba interface (tomba.tumba.pt)

Features and drawbacks

- Search for URL aliases
  - dlib.org, www.dlib.org, dlib.org/index.html
- Required content changes to:
  - reproduce the original layout
  - enable link navigation
- Corrects erroneous media types
  - the correct media type may not be detectable

Archived data

- Harvested from the web and migrated from the tumba! search engine
- Mainly textual contents
- Time span of 4 years (2002-2005)
- 57 million contents,1.5 TB of data

Conclusions

- No criterion alone is enough to define a national web
- The Portuguese web
  - Spread outside the ccTLD
  - Preserving contents from just 3 media types covers most of the web
  - Blogs compose a significant part of a national web
  - Ignoring exclusion mechanisms has little impact on coverage and may be dangerous
- Designing a web user interface for a web archive is a challenging task

Daniel Gomes: dcg@di.fc.ul.pt

tomba.tumba.pt (web archive)

www.tumba.pt (search engine)

xldb.di.fc.ul.pt (research group)

Thank you for your attention.