Digital Libraries and Search Engines, the new Information Systems in the context of Digital Preservation.

Ricardo Campos $^{1,2}$

$^1$ Tomar Polytechnic Institute
Tomar, Portugal

$^2$ Centre of Human Language Technology and Bioinformatics
University of Beira Interior
Covilhã, Portugal

Introduction

Digital Libraries

Implementing a Digital Library

Search Engines

Conclusions
Enterprise Information Systems

Operational Systems

Data Warehouse

Knowledge

Web Farming

Information System
Digital Libraries

Digitization

Web Archiving

Database

Knowledge

Digital Libraries

Introduction
Digital Libraries
Implementing a Digital Library

Search Engines

Conclusions

Information Systems
Internet Growth

Opportunity to transform and readapt the traditional library services

The European Commission predicts that the digitization and digital preservation are activities that will grow up very much in the following years
Search Engines play a fundamental role
Contextualization

Digitization;

Web Harvesting;

Potentialities;

Limitations;

Relevant projects;

Search Engines (preservation, access, and diffusion).
Some initiatives...but fragmented

Actually the most suitable way to interact with a library is by using the Internet, where the user can consult the bibliographic catalogue;

For example, in Portugal, we have:


- The problem is that the initiatives are fragmented;
- The missing of a more general politic constitutes a threat that can implies that some material (physical or digital) get lost;
So, the European Commission promotes in the scope of the Action Plan i2010, the developing of an European Digital Library (http://www.edlproject.eu/) to preserve the memory of Europe (journals, books, etc…);

The project started on September 2006;

The focus of the project, which builds on the existing TEL (The European Library, http://www.theeuropeanlibrary.org/) is to grant access to a set of digital collections by a unique entry point, and should be available around 2008;
In 2004, Google developed Google Books (http://books.google.com), establishing a partnership with a set of libraries (Harvard, Oxford, Stanford, New York Public Library, etc…) with the purpose of digitalize and ease the search for their collections by using Google’s Search Engine;

Performing a simple query the user can obtain basic bibliographic data, a full, limited or snippet preview plus some links to bookstores;

The Open Content Alliance (http://www.opencontentalliance.org/) represents the collaborative efforts of a set of organizations whose goal is to built a digital archive of digitized text and multimedia content for universal access. Content will be soon available through the web site and through Yahoo.
And if we don’t preserve digital contents?

- An increasing exclusive publication of the contents on Internet;
- No prints;

But most of the projects concerns to Digitization

- 60% of the web pages get inaccessible after 1 year;
- 44 days is the average life of a web page;
Who owns the responsibility?

National Digital Libraries;

Research Labs;

Preservation of Digital Culture

Knowledge Protection

Preventing the lost of contents
New perspectives and Opportunities

- Evaluate a website since the beginning;
- Understand the evolution of the web;
- Understand products evolution during a timeline;
- Understand how different people from different countries use the web;
- Consult information of an enterprise or homepage that does not exist anymore;
History

Only recently start appearing a higher number of initiatives based on the web harvesting technique.
Understanding Web Harvesting

- Crawling/Harvesting;
- Parsing;
- Indexing;
- Making it available;
Define a selection criteria?

- Huge quantity of web resources!
  
  Use a selection criteria?

  This has to do with the concept of the project;

  Future consequences on the set of resources available in the future;

2 strategies of web site capture

- Non-selective (Automatic Web Harvesting);

- Selective (Selective Web Harvesting);
Crawling Frequency

It’s obvious that the crawling of e-journals should be the most frequently possible, if needed done daily, trying to capture the highest number of editions;

Other types of publications might just be crawled some times. That should be enough to represent a year of publication;

- Continuously;
- Snapshot;
No matter the choice, the task of doing Web Harvesting it’s not easy;

Although the success of some initiatives, including the famous Internet Archive project, the digital preservation turns hard to do when is done by a unique organization;

That’s the point why some countries start developing new projects
## Web Harvesting Projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Crawling</th>
<th>Accessible?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden, Kulturarw3</td>
<td>A-WH</td>
<td>Limited</td>
</tr>
<tr>
<td>Finland, Eva</td>
<td>A-WH</td>
<td>No</td>
</tr>
<tr>
<td>Denmark, NetArchive</td>
<td>A-WH; S-WH</td>
<td>Limited</td>
</tr>
<tr>
<td>Australia, Pandora</td>
<td>S-WH</td>
<td>Yes</td>
</tr>
<tr>
<td>Canada, e-Collection</td>
<td>S-WH; D</td>
<td>Yes</td>
</tr>
<tr>
<td>Austria, AOLA</td>
<td>A-WH; S-WH</td>
<td>No</td>
</tr>
<tr>
<td>France</td>
<td>A-WH; S-WH; D</td>
<td>No</td>
</tr>
<tr>
<td>Czech Republic, WebArchiv</td>
<td>A-WH; S-WH</td>
<td>Yes</td>
</tr>
<tr>
<td>Lithuania, AER</td>
<td>A-WH; S-WH; D</td>
<td>No</td>
</tr>
<tr>
<td>Netherlands, e-Depot</td>
<td>S-WH</td>
<td>No</td>
</tr>
<tr>
<td>EUA, Minerva and IArchive</td>
<td>S-WH and A-WH; S-WH</td>
<td>Limited and Yes</td>
</tr>
<tr>
<td>Portugal, Tomba</td>
<td>A-WH</td>
<td>Yes</td>
</tr>
</tbody>
</table>
**Approach**

**Automatic Web Harvesting**
- Sweden;
- Finland;
- Portugal;

**Selective Web Harvesting**
- Australia;
- Netherlands;
- EUA (Minerva);

**Combined Approach**
- Denmark;
- Canada;

**Digital Libraries and Search Engines**
- Austria;
- France;
- Lithuania;
- EUA (IA);

---

Digital Libraries and Search Engines, the new Information Systems in the context of Digital Preservation

Ricardo Campos
Access

- Australia;
- Canada;
- Portugal;
- EUA (Internet Archive);
## Software

<table>
<thead>
<tr>
<th>Software</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combine</td>
<td>Sweden; Austria; Finland; Lithuania;</td>
</tr>
<tr>
<td>Heritrix</td>
<td>Denmark; Czech Republic; Australia;</td>
</tr>
<tr>
<td>Tumba</td>
<td>Portugal; EUA (IA); France;</td>
</tr>
<tr>
<td>NedLib</td>
<td>Finland;</td>
</tr>
<tr>
<td>HTTrack</td>
<td>EUA (Minerva);</td>
</tr>
<tr>
<td>Alexa</td>
<td>EUA (IA);</td>
</tr>
<tr>
<td>Xyleme</td>
<td></td>
</tr>
</tbody>
</table>

[www.ipt.pt] [hultig.di.ubi.pt]
Other Projects

- German;
- Japan (WARP);
- Norway (Paradigma);
- England (Domain.uk);
- Greece;
- Europe Archive.
Potentialities

- New ways of access, increasing it’s use;
- Digital Preservation, preventing the lost of contents;
- Readapt the way libraries organize themselves, preventing that the user need’s to go to the library, allowing the consult of contents with restrict access and the simultaneously consult of a document for more than one user;
- Development of technical infra-structures, increasing data storage.
Questions and Limitations

- What to preserve?
- Deep web and Surface web!
- Provide safety and secure access;
- The continuously technical evolution makes it difficult to preserve digital contents;
  
  Up-to-date equipments;

  Guarantee the transfer of digital docs to new supports;

  Maintain the file type integrity as well as the compatibility of the formats.
Questions and Limitations

- The solution:
  
  Maintain the original equipment;
  
  Use actual technology to emulate the original one;
  
  Migration of technical equipment and software to newly one;

- Continuously specialization of human resources;

- Payment method;
Questions and Limitations

- Reorganization of libraries implies some natural expectative but also some fear in librarians;

- Because of it’s natural domain, libraries should take attention for people with special needs, developing services and products easy to use, trying to reach an inclusive information society;

- Copyright!
Defining an on-line access

The development of digital libraries in the context of information systems, is a step forward in the preservation of contents, but it **doesn’t make sense without the definition of an on-line access**;

That should be done by developing an unique entry point:
Search Engines as Information Systems

Search Engines should play a fundamental role in the context of digital libraries, capturing, saving and making the information accessible, they act like an information system;

- Internet;
- Digitization;

Making the information accessible to the user through the development of a graphical interface.
Although it’s importance the relevance parameters defined to retrieve the docs in the context of digital libraries, should be compulsorily different from the ones defined by search engines;

In fact, digital libraries, should not restrict the access to the set of publications more popular, just like PageRank do with webpages;

Aside, search engines suffer from precision and from presenting the information in a structured way;
No matter it’s importance, only 4 web harvesting projects make the contents accessible, but all of them force the user to specify the URL;

Alternative approaches implies the development of web indexes and the definition of keywords capable of describing the resources;

In this context, should be performed an automatic indexing using for that Web Content Mining techniques.
The use of Web Content Mining techniques makes it possible to semantically analyze the docs and to build coherent clusters, easing the search, solving the problem of ambiguous queries and preventing the user need to specify the URL;

In the scope of web search, we have already developed a meta search engine, that can be perfectly adaptable to this context;
WISE System

- Metadata extraction
- Language independent
- Presents the information in a hierarchical structure
- Concept disambiguation
- Progressively Thesaurus development
Opportunities, Difficulties and Approaches

Internet
- Readapt the traditional structure of libraries to a digital structure;

Difficulty
- Consider economical, social, technological and legal aspects;

Approaches
- Digitization;
- Web Archiving;
  - It uses timeline;
  - It can be a support to decision;

[www.ipt.pt] [hultig.di.ubi.pt]
## Reasons, Search Engines

<table>
<thead>
<tr>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation;</td>
</tr>
<tr>
<td>Storage;</td>
</tr>
<tr>
<td>But seldom the collections are accessible;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search Engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Play a fundamental role</td>
</tr>
<tr>
<td>Capture;</td>
</tr>
<tr>
<td>Storing;</td>
</tr>
<tr>
<td>Access.</td>
</tr>
</tbody>
</table>
# Projects, Propose

## Projects
- Tomba;
- Internet Archive;

Need to specify the URL;

To ease the access it’s necessary to build automatic indexes;

## Propose
- Web content mining techniques to semantically analyze and capture the metadata;
- Development of partnerships including the establishment of a framework to save the best practices;
- Study the maturity and evaluation of digital libraries, defining measures that allow to understand the on-line sophistication.
Thanks for your attention!