Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

PORTUGUESE WEB ARCHIVE

Proposal for a project for collaborating with the Portuguese Web Archive

ArqTag: Collaborative tagging of archived content

The FCCN is responsible for the Portuguese Web Archive and is looking to co-operate with Research and Development entities interested in participating in innovative projects. This document presents a proposal for a project with an estimated duration of 1 year, which could for instance be part of the work for a Masters degree.

The Portuguese web is regularly collected and stored for future preservation. This large amount of data requires efficient mechanisms that enable access to the archived information.

Archiving printed publications is subject to a valuable cataloguing process by specialized librarians, which facilitates their later search by the users. Cataloguing Web content by professionals would be valuable to improve search results on archived information, especially for non-textual content, such as images, which is searched by users using keywords.

Unfortunately, the size of the Web does not allow the traditional cataloguing process to be applied in Web archiving. However, it is possible to distribute the task of generating metadata among the user community through the association of descriptive tags to archived content.

It should be noted that professionals like archivists or librarians could continue to provide further contribution by generating higher quality tags, even through formal indexing processes.

Free tagging could seem chaotic compared to the rules used for cataloguing printed publications. On the other hand, as cataloguing rules become obsolete they lose their added value in relation to free tags.

Tags are used to improve Web search engine results. But in Web archives, they have an additional importance because they enable adapting the terminology used in metadata to language evolution over the years.

For example, someone who searched for “War in Iraq” during the 80’s (1980-1988) would probably want to find information about the Iran-Iraq conflict; in 1990 regarding the war caused by Iraq’s invasion of Kuwait; and in 2008 the war caused by the invasion of Iraq by the USA.

Over the years, news published with the title “War in Iraq” was unequivocal at the time they were published and the search engines developed satisfactory responses (assuming they existed in 1980).

In 2010, when someone searches for “War in Iraq”, it is difficult to determine which pages that person is expecting to find. But if the search is refined to “First USA-Iraq War”, referring to the 1990 war, the aim of the search becomes clearer. However, news published during that era didn’t use this terminology. In 1990, journalists didn’t know that a second USA-Iraq war would take place and so the news at that time did not have the terms “First USA-Iraq war”.

A tagging system could help to make these searches less ambiguous because the page metadata is enriched along time by the users. Thus, in 2010 users could tag the three news items about “War in Iraq” in the following way:

- News regarding "War in Iraq" (‘80). Tags: Iran-Iraq War, Ayatollah Khomeini, Islamic fundamentalism.
- News regarding "War in Iraq" (‘90). Tags: First USA-Iraq war, Golf war, USA-Iraq war, Desert storm, Kuwait, oil.
- News regarding "War in Iraq" (‘07). Tags: Second Golf war, Invasion of Iraq, weapons of mass destruction.

If tags were used as a source of information for searches on the archive, the system would then be able to more easily find the desired news for the terms “First USA-Iraq war”.

An archived content may also be tagged with terms in various languages and this information could be used to support multilingual searches on the archived Portuguese Web content.

The objective of this project is to develop a system that enables the Portuguese Web Archive users to tag archived content, similar to what is done through del.icio.us for current web content.

These tags must be stored so that they can be accessed quickly and indexed. Use of RDF or OWL technology or other common techniques from the “Semantic Web” domain can be applied.

The system must include defence mechanisms against malicious users, such as web spammers.

The system could be implemented using existent free open-source software for social networking or from scratch preferably in JAVA.