Proposal for a collaborative project with the Portuguese Web Archive

Guidelines for selection of relevant websites for archive

The Portuguese Foundation for National Scientific Computing (FCCN) is currently engaged in the Portuguese Web Archive (PWA) project and seeks to cooperate with Research and Development organisations who are interested in participating in innovative activities. This document presents a proposal for a project with an estimated duration of 1 year, which could form part of a master's thesis or introduction to research.

The overall objective of the PWA is to regularly compile and archive information from the web relevant to the Portuguese community. The aim is to comprehensively preserve the Portuguese web.

The collection process is performed by a component known as a crawler, which repeatedly compiles, extracts and follows links to new content. The crawler begins its activity from a pool of addresses relevant for archiving, called *seeds*. A proper seed selection is fundamental to ensure the quality of the service provided by the PWA.

Currently, the process of seed selection is based solely on an automatically produced list of website address under the .pt domain. However, there are numerous websites which are of manifest interest to the Portuguese community which don't use the national domain. The automatic selection of websites outside the .pt domain is complex, onerous and has considerable scope for error.

The PWA allows anyone to suggest a website for archiving. However, these suggestions have to be validated before they are included in the list of seeds, according to concise selection criteria. The automatic selection criteria are deterministic, but limited to the interpreting capabilities of machines, while, relevance, as perceived by humans, is subjective.

The main purpose of the proposed work is to investigate criteria for the selection of relevant websites for preservation, resulting in a set of guidelines that would allow a non-expert to identify whether or not a website should be archived. The aim is to validate humanly suggested websites, so that highly relevant content to the Portuguese community is archived. The main product of this project will be a relatively short questionnaire, whose answers will indicate whether or not a given website should be archived.

Bibliography

- Julien Masanès, Web Archiving, 2006.
- Michael Day, Collecting and Preserving the World Wide Web, 2003.
- Daniel Gomes and Mário J. Silva, Characterizing a national community web, ACM Transactions on Internet Technology, 2005.
- Daniel Gomes and Sérgio Freitas and Mário J. Silva, Design and Selection Criteria for a National Web Archive, 2006.
- National Library of Australia, Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia, 2005.