FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

PORTUGUESE
WEB ARCHIVE

Proposal for a collaborative project with the Portuguese Web Archive

Image Search on the Portuguese Web Archive

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.

The web is periodically compiled and archived for preservation purposes. However, little advantage is taken away from this vast wealth of information if there are no efficient search mechanisms. Each content type needs a specialized search service which enables the information to be accessed efficiently. Images are among the most published and searched forms of content on the Web.

The aim of this project is to develop an image search system for the Portuguese Web Archive which enables users to efficiently find and access images compiled over the years. Search on a web archive has characteristics which distinguish it from a conventional search engine. A web archive search system must be able to process larger amounts of data and deal with its temporal features. Current image search engines are based on associated text. However, the association between text and images is not easy to ascertain. The study of efficient mechanisms that identify and associate text with web images is a topic which raises challenging research challenges.

The development of the image search system would be based on the Portuguese Web Archive hardware and software platform. The project would be done in collaboration with the PWA team which presents expertise in web search.

The PWA search system is based on NutchWAX. The Text-based image search capability for NutchWAX (for PWA this version would be required) could form the basis for this project.