Proposal for a collaboration project with the Portuguese Web Archive

Archiving Browsershots of pages during crawling

The Portuguese Foundation for National Scientific Computing (FCCN) is working on the Portuguese Web Archive (PWA) and is looking to co-operate with Research and Development entities interested in carrying out innovative projects. This document presents a proposal for a project with an estimated duration of 1 year.

The PWA regularly collects and stores the Portuguese web. This process is carried out by a component named *crawler*, which repeatedly downloads, extracts and follows links to new content. However, it is necessary to provide the system with mechanisms that allow archived information to be preserved and maintained accessible in the long-term. A problem that web archives face is how to reproduce Web pages, after several years, in a way which is as faithful as possible to the original.

Web archives collect the information available on the Web. However, often the coding of web pages is not compliant with format specifications. Therefore, they can only be correctly rendered by certain browsers versions.

This situation raises two significant problems for web archiving. The first one, is that sometimes the crawler cannot interpret the pages' code and fails to find embedded links and collect content, which may result in archived pages where images are missing. The second problem is that after a few years, it becomes difficult to access archived pages. For example, a few years ago pages were published in a specific code for the Netscape browser. Nowadays, this browser is no longer used and the current browsers have difficulty in correctly presenting the archived pages that contain these specific codes. Apart from that, browsers are developed to respond to characteristics and rules set for the current Web, and concerns about retro-compatibility with specific page codes that only exist in web archives and rarely occur on the current Web are secondary.

Although the pages have code errors which raise problems for crawling and long-term access, normally these pages work with the most common browsers at the time they were available on the Web. The objective of this project is to create an additional component to be included in the crawler that will take "photographs", called browsershots, of the appearance of the pages in the most common browsers, during collection. Thus, even if in the future a browser

cannot correctly interpret the code of an old page to show it correctly, it will always be possible to show an image of its appearance through an old browser at the time it was collected. The problem of preserving pages in the long-term will thereby be mitigated, increasing the possibilities of accessing content across time.

Apart from that, browsershots will make it possible to improve the PWA search system with new functions, such as presenting page thumbnails in search results or quickly browse several versions of a given page across time.

The project will be developed in co-operation the Portuguese Web Archive team, using its hardware, software and data sets. The current PWA crawler uses Heritrix, which is implemented in Java. There are various technologies for generating browsershots, like Screengrab, WebpageDump or Browsershots. Html To Image is used by the Archives Office of Tasmania, integrated with HTTrack. The British library published a list of software to generate browsershots. Generation of browsershots was used in the Zoetrope research work, in which archived web data was processed.