Enhancing access to research the Geocities historical collection

Pedro Gomes and Daniel Gomes, Foundation for Science and Technology: Arquivo.pt
pedro.gomes@fccn.pt, daniel.gomes@fccn.pt

Geocities.com was founded in 1994 as the World Wide Web entered societies worldwide. In 1999, GeoCities was acquired by Yahoo! and it became the third most popular platform due to the ease of creation, management, and dissemination of the websites. However, in April 2009, Yahoo!, announced that it would shut down Geocities, deleting more than 38 million pages created by users all around the world about various subjects. In consequence, the Archive Team announced a project to archive GeoCities pages, resulting in a torrent file which comprises 641 GB of historical information that documents the early days of the Web and how it reflected societies during this disruptive period of Humankind.

The content of the Geocities has been preserved by the Internet Archive or projects such as OoCities, GEOCITIES.ws, and GeocitiesArchive. Due to its historical relevance, the Geocities collection has been used to conduct research studies in several areas such as Arts, Humanities or Sociology. A search by “Geocities” on Google Scholar yielded over 70 000 results since 1994 that included studies focusing on the Geocities historical collection and citing resources that were hosted on the Geocities platform and became unavailable. However, the existing access tools to explore and retrieve information from the Geocities historical collection created by the Archive Team are limited to URL search or browsing.

Arquivo.pt is a research infrastructure that provides access tools over historical web data to support scientific research. It has been developed since 2007 to incrementally improve access to its collections and respond to the needs of researchers. The services provided by Arquivo.pt include full-text search, image search, version history listing, advanced search and application programming interfaces (API) that facilitate the automatic process of large-amounts of historical web data or the development of innovative applications. Arquivo.pt has been used to support research & development activities in several areas by preserving and providing access to valuable scientific resources that became unavailable online and are found by researchers (e.g. past news, data sets, grey literature).

We believe that by improving access to the Geocities historical collection, it has the potential to originate innovative scientific contributions and maintain the scientific value of previous studies that cite Geocities content. Thus, we decided to integrate the Geocities collection in Arquivo.pt so that national and international researchers could benefit from its access tools. This collection can be searched and accessed through the public service available at: arquivo.pt/searchGeocities.

This communication describes the process and challenges of integrating the Geocities historical collection in Arquivo.pt from the torrent file created by the Archive team through an automatic process method and demonstrates the innovative access methods that became available after this integration.

Keywords: Geocities, Web Archive, Digital Humanities, Web History