How to research governmental web data?

Daniel Gomes*[1]

[1]Foundation for Science and Technology (FCT:Arquivo.pt) – FCCN Campus do LNEC Avenida do Brasil, 101 1700-066 Portugal, Portugal

Abstract

The web is the primary means of communication which governments from developed societies use to enable accessibility to open data. For this reason, web pages became the original artefacts that rule modern societies in the Digital Era and the Directive (EU) 2019/1024 of the European Parliament on open data and the re-use of public sector information stipulated that "To facilitate re-use, public sector bodies should, where possible and appropriate, make documents, including those published on websites". On the other hand, the web is a tool for the scrutiny of governmental activities through online news, public discussions or activism platforms. Thus, the web became a crucial resource of information for contemporary political, social or historical research. It is difficult to perform accurate studies about societies during the past 20 years without analysing born-digital information that has been published online.
However, every year, 80% of the pages available online disappear or are updated to a different content. The tremendous fast pace at which the Internet has penetrated modern societies without proper digital preservation makes it difficult to perform research studies that require the analysis of historical web data such as historical web pages published by previous governments or past news. This problem will become even more concerning in the future, when Digital Humanists will look into the past to try to produce a common memory about our present societies and extract knowledge from it.

This communication will describe how a public research infrastructure has preserved web data officially published by governments and at the same time the online information about governmental activities. For this purpose the following research questions were addressed:

> How to select governmental web data to be preserved by combining experts with automated methods (see Bicho & Gomes, Preserving Websites Of Research & Development Projects. In iPRES 2016)?
>
> How to maintain its accessibility across time (see M. Costa "Information search in web archives.", PhD thesis 2014)?
>
> How to provide access to support research over large-scale volumes of historical web data (see https://arquivo.pt/api)?

The Arquivo.pt annual awards were established in 2018 to showcase research works that use preserved historical information from the Web. Studies performed over the historical

*Speaker

web data addressed political, sociological or economical issues. This communication will describe the methods applied in some of these research works to illustrate the utility of the archived web for modern researchers to study web data produced by governments and online news about governmental activities, such as the "Parliamentary Archive" (https://arquivo-parlamento.pt/), Revisionista.pt: Un-cover the news (http://revisionista.pt/), "Economics Archive" (https://arquivo.pt/wayback/20200802200807/https://arquivoeconomico.pt/), "Opinion Archive" (https://arquivo.pt/wayback/20190425145551/http://arquivodeopiniao.pt/en/) or "Politquices" which is a Web application that allows searching support or opposition relations between political personalities and parties expressed in news headlines (https://www.politiquices.pt/). These studies originated open data sets that can also be useful for future works.

Keywords: Digital governance, Arquivo.pt, digital preservation, open data, open access, web archives