How to handle web archive versions?

Niels Brügger*[1], Frédéric Clavert*†[2], Joshgun Sirajzade*[2], and Karin De Wild*[3]

[1]Aarhus University [Aarhus] – Nordre Ringgade 1 DK-8000 Aarhus C, Denmark
[2]University of Luxembourg [Luxembourg] – Campus Kirchberg6, rue Richard Coudenhove-KalergiL-1359 LuxembourgCampus de Limpertsberg162a, avenue de la FaïencerieL-1511 LuxembourgCampus de Belval2, avenue de lÚniversitéL-4365 Esch-sur-Alzette, Luxembourg
[3]University of Leiden – Netherlands

Abstract

The proposed paper will debate different ways of handling versions in a web archive collection. It will present the problem, followed by a discussion on how the existing literature tackles different ways of handling it. Finally, the different approaches will be tested and discussed against the IIPC COVID-19 web archival collection.
For various reasons a collection of archived web material may contain any given web page several times, and often from different points in time. It is an advantage when studying the diachronic development of a web page. It can become a problem when a researcher investigates the entire collection: many distant reading methods may release biased outputs if some web pages of the corpus have different versions, but not all or not along the same modalities. Therefore, in studies that investigate an entire collection researchers would prefer to have only one copy of each entity. Hence a fundamental question: which version of a webpage should they choose?

The answer to this question is even more complex, that there are two ways of being 'the same': copies – two web pages are identical in any respect – and versions – they share the same URL, but they have been archived at different points in time and they are not exactly identical (Brügger, 2018). This paper will focus on versions: as versions may come with varying degrees of being non-identical, from a single character having been changed to an entirely new web page with the same URL: versions emerge at the nexus of variation and temporality.

In the existing literature different approaches have been used to deal with versions: for instance, some have set a specific point in time as the yardstick against which copies or versions should be evaluated to support the choice of only one of each (Weltevrede, Helmond, 2012; Helmond, 2017; Fage-Butler et al., 2021), while others have focused on providing the most complete version (Brügger et al., 2020). The overall rule that is at stake here is one of balancing temporal consistency against completeness.

We will discuss and test several strategies to deal with versions against the IIPC Covid-19 corpus. That will allow us to evaluate different strategies in terms of balance between temporal consistency and completeness.

Literature

*Speaker
†Corresponding author: frederic.clavert@uni.lu

Brügger, N. (2018). The archived web: Doing history in the digital age. Cambridge, MA: MIT Press.

Brügger, N., Nielsen, J., Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. First Monday, 25(3), https://firstmonday.org/ojs/index.php/fm/article/view/10384.

Fage-Butler, A., Ledderer. L., Brügger, N. (2021, unpublished). Proposing methods to explore the evolution of the concept 'mHealth' on the Danish Web archive.

Helmond, A. (2017). Historical website ecology: Analyzing past states of the web using archived source code (pp. xx-xx). In N. Brügger (Ed.), Web 25: Histories from the First 25 Years of the World Wide Web. New York: Peter Lang Publishing.

Weltevrede, E., Helmond, A. (2012). Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. First Monday, 17(2), https://firstmonday.org/ojs/index.php/fm/article/view/377

Keywords: version, duplicate, web archive