**Unlocking web archives through metadata, seed lists and derived data**

Frédéric Clavert and Valérie Schafer (C2DH, University of Luxembourg)

This proposal aims to address the use, re-use, access and dissemination of data related to web archives. Web archives (Brügger, 2018) have been for several years in a hybrid position regarding access, depending on the institutions that were preserving them. While Internet Archive has made its collections available online since 2001 through the Wayback Machine (but with limited features for scholars willing to conduct a distant reading based on data, WARC files, etc.), most national libraries only allowed an onsite access due to authors rights restrictions (and in some cases the frame of legal deposits), while starting to provide interesting metadata for research projects willing to explore them.

However, the situation is currently evolving in the frame of several research projects that allow to access a vast amount of (international) metadata and datasets. Taking two research projects in progress as case studies, WARCnet and AWAC2, this paper aims to present the move towards the use of metadata and derived data related to huge collections of web archives of the COVID crisis.

WARCnet (Web ARChive studies network researching web domains and events) is a network whose activities (funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B)) run in 2020-2023. The networking activities are guided by overarching research questions, one of them being “How transnational events developed on the European web?” (and notably the COVID crisis which is explored in WG2 (<https://cc.au.dk/en/warcnet/working-groups)>).

AWAC2 (Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset) is a project part of the [Archives Unleashed Cohort Program](https://archivesunleashed.org/cohorts2021-2022/), that supports and facilitates research engagement with web archives. It aims to explore a unique collection of web material (<https://archive-it.org/collections/13529>) related to the pandemic, with contributions from over 30 members of IIPC (International Internet Preservation Consortium) as well as public nominations from over 100 individuals/institutions.

May it be in terms of access or tools, both projects are currently exploring new methodologies based on broad datasets (i.e. 5,3 TB for the IIPC collection related to the COVID crisis; 9.4 GB and 8,738,751 lines for the CSV related to plain text webpages).

Starting with the WARCnet project, the presentation will explain how its WG2 gathered and accessed several national European datasets of COVID web archives, their specificities as well as their heterogeneity, the first analysis conducted through a datathon on January- February 2021 (Aasman et al. 2021) and the limits and assets of such access.

Within the AWAC2 project (2021-2022) the access to the international IIPC COVID collection, through Archive-It and through the cohort program developed by the Archives Unleashed Team (Netpreserve, 2021; Ruest et al., 2021), is then a new opportunity to access data through mediated interfaces (ARCH) and to go further into them. Here again the presentation will demonstrate new opportunities and show a few examples of the analysis conducted by the team.

Both examples aim to present the way web archiving institutions, libraries and researchers are developing new ways of accessing and exploring web archives, while also increasing their value(s) (Schafer and Winters, 2021).

**References**

Aasman, S., Bingham, N., Brügger, N., de Wild, K., Gebeil S. & Schafer V. (2021). Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID- 19 Special Collections, *WARCnet paper*, Aarhus,

https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman\_et\_al\_Chicken\_and\_Egg.pdf

Brügger, N. (2018). *The Archived Web. Doing History in the Digital Age*. Cambridge, MA: The MIT Press.

IIPC (2021), A Retrospective with the Archives Unleashed Project, netpreserve blog, https://netpreserveblog.wordpress.com/2021/04/01/a-retrospective-with-the-archives-unleashed-project/

Ruest, N., Fritz, S., Deschamps, R. Lin, J. & Milligan, I. (2021) From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*, https://paperity.org/p/260049927/from-archive-to-analysis-accessing-web-archives-at-scale-through-a-cloud-based-interface

Schafer V. & Winters J. (2021). The values of web archives, *International Journal of Digital Humanities,* 1-10, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8190571/