Continuity and discontinuity in web archives

Quentin Lobbé*[1]

[1]Complex Systems Institute of Paris île de france (ISCPIF CNRS) – CNRS : URA3611 – France

Abstract

Memory institutions have been archiving the Web for the last 25 years. These initiatives seek to preserve
pieces of our digital heritage by harvesting web resources. But just like Funes in Borges short story, memory institutions
will never be able to achieve exhaustiveness. Archiving always goes with complex selection criteria stated by librarians,
curators, politicians or – in the particular case of web archiving – engineers and robots called crawlers that determine the
spatio-temporal coverage of archived corpora. Unlike traditional archival materials, web archives can't be understood
apart from their own archiving processes : crawlers tear web resources away from the continuous temporality of the
Web and produce discretized snapshots timestamped by archiving date. By nature, Web archives are not direct traces
of the Web, they are direct traces of crawlers (1).
The web archives movement has originally been sparked with the intuition that web resources were intended to
become valuable research materials in the hands of future historians ; and archiving pioneers indisputably succeeded
in preserving what could be saved in a short period of time. Yet, the exploration of available archives corpora remains
an open problem. Convincing qualitative approaches do exist, but exploration tasks quickly become non-trivial when
scholars widen the scope of their analysis. At a macro scale, archives collections suffer from being too wide, too rich
and paradoxically often incomplete : they lack spatio-temporal continuity. We here think that the key of future large
scale analysis will depend of our capacity to resolve the problem of discontinuity induced by the inner nature of web
archives. We aim for re-injecting continuity in web archives.
In this article, we propose a theoretical and practical framework for reconstructing continuity spaces inside web
archives corpora. Continuity spaces are multi-level evolving sets of coherent archived web resources within which
researchers can conduct historical analysis at scale. We illustrate our approach by studying the history of the firsttuesday
community : a dead constellation of networking web sites that acted in the interest of the

*Speaker

economical growth of the
Web before the bursting of the dot-com bubble in the early 2000's.
Our contribution has practical value for web archives explorers (historians, sociologists, anthropologists, etc.) aiming for conducting large scale analysis of the past Web ; that is, any diachronic study that goes beyond purely qualitative strategies. For our's part, we reserve the use of web archives for a more systemic purpose. In line with Complex
Systems approaches, think that our work could be a first stone in the study of the morphogenesis of the Web. We thus
consider the Web as a complex and dynamic ecosystem shaped by networks of evolving processes (technical, economical, sociological, political, etc.) where entities interact with each other (web sites, users, robots, opinions, etc.) at
various levels of organization. By solving this problem of continuity and discontinuity, we could eventually use web
archives as raw materials for the phenomenological reconstruction of the genesis and evolution of the Web from the
mid 90's to the present days.
In order to comprehend and solve the problem of continuity and discontinuity in web archives, we choose to address
two concrete research questions :
1. How can we reconstruct the evolution of a web site / group of web sites from its / their full set of archived traces ?
2. How can we reconstruct the evolution of an historical event at scale whose traces can be found in web archives ?
These questions complement one another as they investigate web archives from the angle of both containers (1.)
and content (2.).
Références
(1) Lobbé, Q. (2018, November). Where the dead blogs are. In International Conference on Asian Digital Libraries (pp. 112-123). Springer, Cham.

Keywords: continuity, discontinuity, multi, level explorations, web fragments, phylomemy