Understanding Web Archives

Helen Hockx-Yu

Head of Web Archiving
British Library, London, UK,
Helen.Hockx-Yu@bl.uk

Abstract. This talk provides an insight into web archives by examining the "unknown" aspects beyond the archived web pages, or the "text". It argues that web archives have a rich set of semantics which when explored offers a new way of understanding their characteristics. It showcases examples of British Library's work beyond the "document-centric" approach of providing access.

Keywords: Web archives, exploration, semantics

1 Introduction

The effort to archive the web started in the mid-1990s, a few years after the web was born. This was initiated by the Internet Archive in the US. Many national libraries and archives, which traditionally have the duty to preserve a nation's cultural and scientific heritage, followed the suite and started actively collecting web content. Internet Archive's Wayback Machine[1] is the earliest and most comprehensive web archive to date, containing over 435 billion web pages archived from 1996. Many national heritage organisations have established collections covering their respective national web domain or subsets of it.

There are however issues related to the access and use of web archives: it is often restricted by legal requirements on one hand, in exchange for reproducing copyrighted material for the purpose of cultural heritage, and by the (single) envisaged use case on the other [HY14]. The latter is based on the assumption of web archives consisting of historical documents (web pages) used for reference. Researchers access previous states of individual web pages and websites in a web archive, which are selected, described and grouped together by curators, in the same way as printed books and journals. The over-focus on "documents" or "text" means contexts of archived material tend to be ignored or regarded as irrelevant.

2 Understanding Web Archives

A common assumption is that web archives contain copies of older versions of websites which are no longer current and have been replaced by the "live"

[1] https://archive.org/web/web.php

version. Brügger and Finnemann argue that archived web resources are "reborn", different from digitised and born digital collections and from the live web in many ways [BF13]. Using the British Library's web archive as an example, this talk examines in detail the many boundaries and imitations related to web archive, determined by purpose, strategy, legal requirements and technological choices. It also points out a fundamental oversight which impacts users' interpretation or understanding of web archives: very little is explained or made clear to the users beyond the actual HTML pages (or the "text"). A typical example of this is the common error message "Resource Not in Archive", which is presented to the end-users when a requested URL cannot be found in the archive. This could be caused by many reasons: some are intended, introduced by things like data limitation at crawl time or content beyond the scope of the crawl; others relate to technical limitations, e.g. dynamic content which the web crawlers are not capable of collecting.

3 More to "text"

Effort started to emerge in recently years which moves away from the level of single webpages or websites to the entire web archive collection. Using visualisation and data analytic techniques, new ways have been developed to view web archives, offering opportunity to unlock embedded patterns and trends, relationships and contexts, which are not possible by consulting websites individually. This is in alignment with the changes in scholarly practices as researchers increasingly take advantage of new possibilities offered by technology. New methods of scholarship are emerging, which challenges the primacy of "text" as object of study. This talk references the concepts of "paratexts"[Nie10] and "distant reading"[Mor00], as theoretical basis for using web archives as scholarly sources. The role of web archives is to provide services supporting scholars who read texts differently.

This talk focuses on a range of non-text attributes of web archives (including an example visualisation or demo for each), explored by the British Library or others, as additional ways of understanding web archives. Scholars are encouraged to explore other contextual or "para-textual" content in the web archives, such as viral content and crawl logs.

- Statistical overview, scale and distribution of a web national domain
- Size: bytes
- Space: geo location, postcodes
- Type of content, e.g. file format, language
- Structure, linked entities and networks
- Evolution, pattern of change over time, e.g. domain names
- Correlation, e.g. between certain term and historical event

This talk also discusses the general issues related to analytical access, such as researchers' scepticism or suspicion about hidden algorithms behind analysis, and how biases in data and how data collection decisions lead to variances in outputs.

References

[BF13] Niels Brügger and Niels Ole Finnemann. The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media*, 57(1):66–80, 2013.

[HY14] Helen Hockx-Yu. Access and scholarly use of web archives. *Alexandria*, 25(1):113–127, August 2014.

[Mor00] Franco Moretti. Conjectures on world literature. `http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature` (accessed on 17 November 2014), 2000.

[Nie10] Niels Brügger. Website analysis: Elements of a conceptual architecture. `http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/cfis_skriftserie/012_brugger.pdf` (accessed on 17 November 2014), 2010.