Analysis of Cyber Threats Affecting the Survivability of Online Digital Projects

Meneses, Luis
luis.meneses@viu.ca
Vancouver Island University

Martin, Jonathan
jonathan.d.martin@kcl.ac.uk
King's College London

Fitzpatrick (Fitzpatrick 2011) focuses on the technological changes (notably greater utilization of Internet publication technologies and digital archiving) that are necessary to allow academic publishing to thrive into the future, highlighting that the key issues that must be addressed are social and institutional in origin. Along these lines, our research incorporates machine learning and data mining to address digital preservation questions, while focusing on identifying identify and managing the signs of degradation of online projects in the digital humanities (Meneses et al. 2020).

The degradation of online projects is not a binary problem, but one that occurs gradually. While recommendations are in place for creating stable and long-lasting resources (Arneil / Holmes / Newton 2019), online projects will show signs of degradation if they don't receive updates, both in their content and in their underlying operating systems. "Bit rot" is a term used in digital archiving to describe the way digital files can spontaneously and quickly decompose. "Link rot" is more specific to the degradation of online projects: over time hyperlinks can cease to point to their originally target due to that resource being relocated to a new address or becoming permanently unavailable (Spinellis 2003); which makes Web pages ephemeral (Markwell / Brooks 2002). However, identifying the gradual decomposition of online digital projects is a more nuanced problem.

For this purpose, we have created a dataset based on the URLs for the companion websites that are referenced in the Book of Abstracts published by the Alliance of Digital Humanities Organizations from 2006 to 2022. Using this dataset, we have defined and identified the shelf life of its companion websites. The *shelf life is* the average length of time that a companion website can endure without updates until it can ultimately be considered abandoned by its researcher, which we have approximated to 5 years (Meneses / Furuta).

This analysis has been a continuous process, in which we periodically create a set of WARC files for each online resource, which are processed using Python (van Rossum 1995) and Apache Spark (Apache Software Foundation 2017) to statistically analyze the retrieved HTTP response codes, number of redirects, DNS metadata and detailed examination of the contents and links returned by traversing the base node. This combination of metrics and techniques has allowed us to identify the most important signature for the degradation of a website in the dataset: the validity and overall health of the topology of links (Meneses et al. 2019). However, this analysis only paints a partial picture in assessing the degree of change of a website over time: the security of the hosts from where these online projects are served is an important factor that should be considered.

We propose to leverage public vulnerability scanners (e.g. shodan.io) to examine the servers where online digital projects are hosted and list their common vulnerabilities. Provided in an anonymized, aggregated fashion, this assessment will shed light on our hypothesis: cyber threats are a significant factor in the degradation of online projects. These threats which might be malicious hacking, ransomware, or other bot-driven disruptions, are prevalent on the public Internet. With online projects increasingly turning to commercial service providers (quantified and mapped in our study), typically providing blank-slate unmanaged virtual servers with limited server administration and threat detection, the need for an assessment of the potential vulnerabilities of digital online projects becomes important. Our results will show numerous vulnerabilities that directly threaten the survivability of online digital projects.

Another problematic case that we propose to explore is when hosts serving online projects are also engaged in malicious activities (without the awareness of principal researchers). For this, we are extending our efforts by including the ADHO data and the resources referenced in Digital Humanities Quarterly. We will use public threat feeds (e.g. greynoise.io) to explore these cases. This analysis will help to further underscore the need for technical expertise to ensure the long-term survivability of online projects, and, as such, the intellectual history of the digital humanities. In the end, we intend this study to contribute to the efforts of the community towards the better preservation strategies to ensure the sustainability and viability of projects in the digital humanities.

Bibliography

Apache Software Foundation (2017): *Apache Spark: Lightning-fast cluster computing*. http://spark.apache.org [letzter Zugriff 11. April 2017].

Arneil, Stewart / **Holmes, Martin** / **Newton, Greg** (2019): *Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH*. Utrecht, The Netherlands.

Fitzpatrick, Kathleen (2011): *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: NYU Press.

Markwell, John / **Brooks, David W.** (2002): "Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks", in: *Journal of Science Education and Technology* 11 (2): 105–108. 10.1023/a:1014627511641.

Meneses, Luis / **Furuta, Richard**: "Shelf life: Identifying the abandonment of online digital humanities projects", in: *Digital Scholarship in the Humanities*: 10.1093/llc/fqy079.

Meneses, Luis / **Martin, Jonathan** / **Furuta, Richard** / **Siemens, Ray** (2019): *A Framework to Quantify the Signs of Abandonment in Online Digital Humanities Projects*. Utrecht, The Netherlands.

Meneses, Luis / **Martin, Jonathan** / **Furuta, Richard** / **Siemens, Ray** (2020): *Analyzing Link Topology to Quantify the Degree of Planned Obsolesce in Online Digital Humanities Projects*. Ottawa, ON, Canada.

Spinellis, Diomidis (2003): "The decay and failures of web references", in: *Communications of the ACM* 46 (1): 71–77. 10.1145/602421.602422.

van Rossum, Guido (1995): *Python tutorial, Technical Report CS-R9526*. Amsterdam: Centrum voor Wikunde en Informatica (CWI). https://ir.cwi.nl/pub/5007/05007D.pdf.