PART DEUX: EXPLORING THE SIGNS OF ABANDONMENT OF ONLINE DIGITAL HUMANITIES PROJECTS

21 Jun 2018 PART DEUX: EXPLORING THE SIGNS OF ABANDONMENT OF ONLINE DIGITAL HUMANITIES PROJECTS

Luis Meneses (ldmm@uvic.ca), Electronic Textual Cultures Laboratory – University of Victoria, Canada and Jonathan Martin (jonathan.d.martin@kcl.ac.uk), King's College London and Richard Furuta (furuta@cse.tamu.edu), Center for the Study of Digital Libraries, Texas A&M University and Ray Siemens (siemens@uvic.ca), Electronic Textual Cultures Laboratory – University of Victoria, Canada

XML (wp-content/uploads/2018/05/MENESES_Luis_Part_Deux__Exploring_the_Signs_of_Abandonment_o.xml)

1. INTRODUCTION

Building online research components for projects in the digital humanities is a

common practice. However, not many researchers have a plan for these online components once the project halts or comes to an end. Consequently, many of these projects become abandoned and slowly degrade over time –some more gracefully than others. Additionally, there is a certain inherent fragility associated with software and our online research tools. In turn, this fragility threatens the completeness and the sustainability of our work over time.

Previous studies have attempted to harness and manage the fragility of online resources. Studies have been carried out to address their potential reconstruction (Klein et al., 2011), the overall decay of websites (Bar-Yossef et al., 2004) and the decomposition of their shared resources (SalahEldeen and Nelson, 2012). Recently, our research has been focusing on analyzing the perceptions of change in distributed collections (Meneses et al., 2016). However, we believe that the inherent characteristics of online digital humanities projects present an interesting (and unique) area for inquiry for two reasons. First, the research aspect of digital humanities projects hinders previous approaches –as the methods for identifying change in the Web do not fully apply. And second, digital humanities projects have a limited useful life –which is accompanied by research from primary investigator, which may or may not be indicated by updates in the project's content and tools.

We presented a paper in Digital Humanities 2017 that explored the abandonment and the average lifespan of online projects in the digital humanities (Meneses and Furuta, 2017). However, we believe that managing and characterizing the online degradation of digital humanities projects is a complex problem that demands further analysis. In this abstract, we propose to explore further the distinctive signs of abandonment of online digital humanities projects. For this second instalment of our study we took a different direction: we departed from strictly using retrieved HTTP response codes and incorporated additional metrics such as number of redirects, DNS metadata and a detailed analysis of content features.

This study aims to answer four questions. First, can we identify abandoned projects using computational methods? Second, can the degree of abandonment be quantified? Third, what features are more relevant than others when identifying instances of abandonment? Our final question is philosophical: can an abandoned project still be considered a digital humanities project?

2. METHODOLOGY

A complete listing of research projects in the Digital Humanities does not exist. However, the Alliance of Digital Humanities Organizations publishes a Book of Abstracts after each Digital Humanities conference as a PDF. Each one of these volumes can be treated as a compendium of the research that is carried out in the field. To create a dataset, we downloaded the Books of Abstracts corresponding from 2006 to 2016 –except for 2015 which was not available for download. We must thank and acknowledge Dr. Jason Ensor from Western Sidney University for providing us the abstracts for the 2015 Digital Humanities conference –which completes our dataset of conference abstracts. We obtained these abstracts after we had carried out our preliminary analysis. Therefore, we will present our findings using the complete dataset of abstracts in the presentation of our paper.

Then we proceeded to extract the text from these documents using Apache Tika and parse the 5845 unique URLs that we found using regular expressions. Then we used Python's Requests Library to retrieve the HTTP response codes and headers corresponding to the URLs, which we used to classify the websites into two groups depending on their correctness: valid (correct) and decayed (showing signs of degradation). Figure 1 shows the distribution of decay for each year. Based on our preliminary findings we approximate the average lifespan of a research project to 5 years, which aligns with reports from previous work (Goh and Ng, 2007). The average time in years since the last modification of the websites in the study is shown in figure 2.

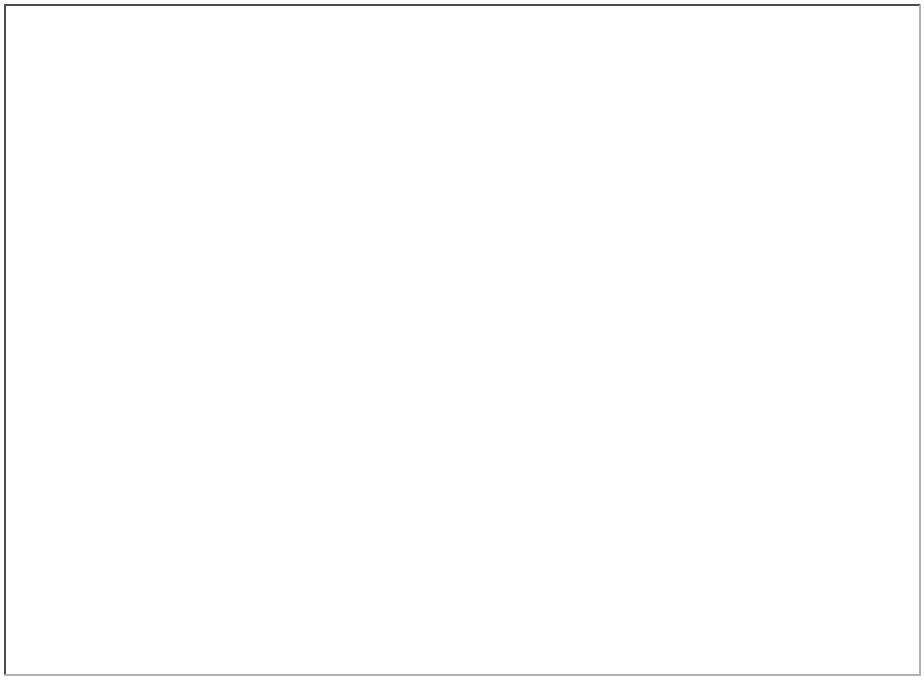

Figure 1: URL decay by year.

Figure 2: Average time in years since last modification.

3. DEVELOPING CLASSIFIERS

To develop classifiers for the degradation identified in the previous section, we considered features computed based on DNS metadata, the initial HTTP request, number of redirects, and the contents and links returned by traversing the base node. The features we included are divided into topology, content-type, anchor-text and child node features. These features stem from concepts we used in our previous work (Meneses et al., 2016).

The text associated with resources is the most obvious feature for determining the topics. Given that we are dealing with a very specialized domain, we developed a domain-oriented expectation model. In particular, we generated topic and term frequency models to examine the similarity among the documents in a given project (the contents of the base node and the metadata and the contents of the child nodes). We used Latent Dirichlet Allocation to model the content of the text (Blei et al., 2003)
and a simple Tf-Idf ranking function to measure and compare them. This ranking function is based on adding the Tf-Idf values for the documents, which were calculated using the terms from the topic modelling as a vocabulary. We will present a detailed version of our results on the longer version of our paper.

4. DISCUSSION

This study is an attempt to categorize change in a very specific domain. More so, this study constitutes one step towards addressing potential strategies for the archival and the long-term preservation of abandoned digital projects. It is important to highlight that not all projects are equal and thus require different approaches towards long-term preservation. In the case of dynamically generated projects, a common approach nowadays is to produce a static set of HTML files which are easier to store. However, this approach assumes the backwards compatibility of Web browsers over time –something that has not always been the case.

To summarize, in this study we aim to computationally identify the indicators of the abandonment of digital humanities projects –as well as quantify their degrees of neglect. To address our philosophical question, we believe that an abandoned project can still be considered a valid digital humanities project depending on its audience. However, this has several nuances that should be considered. Digital online projects in the humanities have unique characteristics that make them impervious to the metrics that used in the Web as a whole –which make them worthy of study. In the end, we intend this study to be a step forward towards better preservation strategies and for the planned obsolesce of digital humanities projects.

APPENDIX A

Bibliography

1. **Bar-Yossef, Z., Broder, A. Z., Kumar, R. and Tomkins, A.** (2004). Sic transit gloria telae: towards an understanding of the web's decay. ACM doi:10.1145/988672.988716.
2. **Blei, D. M., Ng, A. Y. and Jordan, M. I.** (2003). Latent dirichlet allocation.
   The Journal of Machine Learning Research,
   3: 993–1022.
3. **Goh, D. H. and Ng, P. K.** (2007). Link decay in leading information science journals.
   Journal of the American Society for Information Science and Technology,
   58(1): 15–24.
4. **Klein, M., Ware, J. and Nelson, M. L.** (2011). Rediscovering missing web pages using link neighborhood lexical signatures. ACM doi:10.1145/1998076.1998101.

5. **Meneses, L. and Furuta, R.** (2017). Shelf life: Identifying the Abandonment of Online Digital Humanities Projects Paper presented at the Digital Humanities 2017, Montreal, Canada.
6. **Meneses, L., Jayarathna, S., Furuta, R. and Shipman, F.** (2016). Analyzing the Perceptions of Change in a Distributed Collection of Web Documents.
Proceedings of the 27th ACM Conference on Hypertext and Social Media. (HT '16). New York, NY, USA: ACM, pp. 273–278 doi:10.1145/2914586.2914628. http://doi.acm.org/10.1145/2914586.2914628 (accessed 12 April 2017).
7. **SalahEldeen, H. M. and Nelson, M. L.** (2012). Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?.

« Unsustainable Digital Cultural Collections (https://dh2018.adho.org/en/unsustainable-digital-cultural-collections/)

devochdelia: el Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas Indíjenas Americanas de Rodolfo Lenz en versión digital »

LEAVE A COMMENT

You must be logged in (https://dh2018.adho.org/wp-login.php?redirect_to=https%3A%2F%2Fdh2018.adho.org%2Fen%2Fpart-deux-exploring-the-signs-of-abandonment-of-online-digital-humanities-projects%2F) to post a comment.

EL COLEGIO DE MÉXICO (https://www.colmex.mx)

Universidad Nacional Autónoma de México (http://www.humanidadesdigitales.net/)

RED DE HUMANIDADES DIGITALES (http://www.humanidadesdigitales.net/)

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS (https://www.adho.org)

LANGUAGE

English (EN)

Sitio y video desarrollado gracias al Programa de Educación Digital (PRED) de El Colegio de México