Finding Community in the Ruins of GeoCities: Distantly Reading a Web Archive

Ian Milligan Department of History University of Waterloo 200 University Ave. W, Waterloo ON N2L 3G1 Canada i2millig@uwaterloo.ca

ABSTRACT

This paper provides a brief overview of my work with the GeoCities web archive. Asking the question of "can we find community," I use it as a case study to explain various methods for distantly reading web archives.

1. INTRODUCTION

In the early 1990s, as the Web became part of everyday life, GeoCities provided a sense of community to facilitate users' first, often hesitant, steps into the world of Web publishing. At that time, GeoCities was one of the largest collections of noncommercialized, user-generated, public speech ever assembled. Along the digital "GeoAvenues" and in the provocatively-named "homesteads," users developed a relationship with the Web that built a foundation for the blogging and social networking explosion of the 2000s. Rather than simply being a platform upon which users created individualized webpages, GeoCities users helped each other develop the site as a whole.

I explore how we can find community in the ruins of GeoCities. Using a combination of distant, computational reading (studying websites as a collective whole, rather than as individual documents) and more focused, targeted reading, I argue that we can understand the GeoCities web archive as ruins of a vibrant, interconnected community.[1]

Overcoming these technical challenges is of critical importance to the historiography on the early web, however. In the digital ruins of GeoCities, new users figured out their relationship to the Web. They were not alone; they were part of a larger community. As so much of our relationship with web services is now mediated by user interfaces and glossy tutorials, it is worthwhile to look back to the early 1990s and how -- spread out across time and space -- users figured out what the Web would mean to them. GeoCities, a massive assemblage of noncommercialized public speech, presents an interesting way to explore the history of the early World Wide Web.

2. IMPORTANCE

Big Data is coming to history. The advent of web-archived material from 1996 onwards presents a challenge. In my work, I explore what tools, methods, and approaches historians need to adopt to study web archives.

GeoCities lets us test this. It will be one of the largest records of the lives of non-elite people *ever*. The Old Bailey Online can rightfully describe their 197,000 trials as the "largest body of texts detailing the lives of non-elite people ever published" between 1674 and 1913.[2] But GeoCities, drawing on the material we have between 1996 and 2009, has over thirty-eight million pages.



GEOCITIES USERS:

Figure 1. Rapid growth of an online community.

These are the records of everyday people who published on the Web, reaching audiences far bigger than previously imaginable.

GeoCities was a unique place. It rapidly grew, reaching millions of users in a few years. It concept helped to bridge the locally-based networks of the early 1990s and BBSes with the wide open Web - it clustered users together in neighbourhoods, from the children-focused EnchantedForest, the family-focused Heartland, the education-focused Athens. Users relied on each other to find content: from living next to each other in neighbourhoods, to linking to each other using Web rings. In an era before the widespread adoption of search engines, which often required explicit SEO techniques and effort to be part of, these were important.

This unique early experiment in the history of the Web came to halt in 2009 when Yahoo! shuttered it (after acquiring it in a major 1999 deal).[3] If it hadn't been for the timely intervention of Archive Team and others such as the Internet Archive, GeoCities would have been lost forever.[4]

It would be as if the Old Bailey had been thrown on the fire pit of history.

Ian Milligan. "Finding Community in the Ruins of GeoCities: Distantly Reading a Web Archive." Bulletin of IEEE Technical Committee on Digital Libraries, Vol. 11, Issue. 2 (October 2015). This article is distributed under the terms of the Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

3. WHAT CAN WE DO?

To gain access to the GeoCities collection, one has to download the Archive Team torrent or use the Wayback Machine. The latter is accessible at <u>http://archive.org/wayback</u>. If you know particular URLs of webpages that you want to look at, it is very useful; however, if you want to systematically explore Geocities, the torrent is key.

The torrent is a large download (641GB), and is also available at the Internet Archive as a normal download (beginning with <u>https://archive.org/details/2009-archiveteam-geocities-part1</u>). As it is not in WARC format, it presents challenges, as normal web archiving workflows do not work. One ends up with a mirrored web structure on their hard drive, requiring constant workarounds to generate meaningful derivative data.

However, through text analysis and data mining, we can begin to approach the question of whether GeoCities users experienced community, and whether the neighbourhood system worked. We can do this through a few quick ways.

We can:

- use topic modeling to take various neighbourhoods and communities and see what recurring concepts appear; do these line up with the descriptions? Was there coherence? Preliminary results indicate that "neighbourhoods" largely lined up with the intended descriptions, although many developed more narrow focus. For example, the largest community "Heartland" became more focused on Christianvalues and genealogy over time;



Figure 2 – Word Cloud of "community leaders" pages. An imperfect visualization that gives a sense of their interests.

- find the 'community leaders' that provided instructional support, extract their pages, and run content analysis to see what they provided. Word frequency and close reading indicates that most of the instruction of HTML and web site construction came from users themselves, who offered help to targeted subsets of users;
- move between distant reading and close reading, using link structures and topics, to tell detailed stories. In an era before widespread search, important nodes emerge that connected many websites: they facilitated user discovery. Using network analysis, specifically in-degree calculations, we can find these hubs; and
- extract hundreds of thousands of images, arrange them, and begin to see different profiles that tie neighbourhoods together.

My findings suggest that the web archives of GeoCities reveal a vibrant, interconnected community of users taking their first steps out on the Web. It makes a case for why this sort of work matters, both to our understanding of the early Web as well as human culture more generally in the 1990s.

Web archives are the archives of the future, and to access them requires the use of robust digital methods. Historians need to be ready for this impending shift.

4. NEXT STEPS

My next steps are to work with the Web ARChive (WARC) files themselves, which are held by the Internet Archive. I believe this would be a fantastic test corpus, as it is cleaner than other promising datasets such as the 1996-2000 end-of-millenium crawl collected by the Internet Archive.

I would like to incorporate this into the Warcbase platform, which I have been working on with Jimmy Lin (Waterloo)[5]. We have a series of steps that bring you from a directory of WARCs to all the standard scholarly derivatives that informed this work, as well as others. Right now it is too difficult for researchers to do this work on their own.

5. ACKNOWLEDGMENTS

My sincerest thanks to the Social Sciences and Humanities Research Council of Canada (SSHRC), which funded this project through an Insight Development Grant and an Insight Grant. Student support has been forthcoming from the Ontario Ministry of Research and Innovation's Early Researcher Award program. The University of Waterloo has also supported this work.

6. REFERENCES

- [1] A more developed version of this paper "Welcome to the Web: The Online Community of GeoCities and the Early Years of the World Wide Web" is slated to appear in a forthcoming anthology on early Web history in 2016. It will contain more information on the history of GeoCities, go into detail on each of the findings, but contain very little technical information.
- Old Bailey Proceedings Online, "Old Bailey Online The Proceedings of the Old Bailey, 1674-1913 - Central Criminal Court," 2013. [Online]. Available: http://www.oldbaileyonline.org/. [Accessed: 12-Mar-2015].
- [3] J. Motavalli, Bamboozled at the Revolution: How Big Media Lost Billions in the Battle for the Internet. New York, Penguin, ISBN 978-0142002896, 2004.
- [4] Internet Archive, *The Archive Team Geocities Snapshot* (*Part 1 of 8*): *Free Download & Streaming*: *Internet Archive*. 2009.
- [5] "lintool/warcbase," *GitHub.* [Online]. Available: https://github.com/lintool/warcbase. [Accessed: 11-May-2015].