data

MDPI

Data Descriptor

Four Datasets Derived from an Archive of Personal Homepages (1995–2009)

Sean C. Rife

Department of Psychology, Murray State University, Murray, KY 42071, USA; srife1@murraystate.edu; Tel.: +1-270-809-4404

Academic Editor: Xinyue Ye
Received: 31 December 2016; Accepted: 8 June 2017; Published: 13 June 2017

Abstract: While data from social media are easily accessible, understanding how individuals expressed themselves on the Internet in its initial years of public availability (the mid-late 1990s) has proved difficult. In this data deposit, I describe how archival data from Geocities homepages were retrieved and processed to remove non-text data, then further refined to create separate datasets, each of which provides unique insights into modes of personal expression on the early Internet. The present paper describes four datasets, all of which were derived from a larger collection of personal websites: (1) a large corpus of raw text data from Geocities personal homepages, (2) a linguistic analysis of basic psychological properties of the same Geocities pages, using an open-source implementation of the Linguistic Inquiry Word Count (LIWC), (3) a dataset of links between homepages (suitable for network analysis), and (4) a manifest dataset summarizing the size and last update date for each file in the dataset. Data from over 378,000 Geocities pages are included. In addition to providing a detailed description of how these datasets were created, I describe how they might be utilized in future research.

Data Set: https://osf.io/8u48d/

Keywords: Internet; linguistics; online culture; Linguistic Inquiry Word Count (LIWC); corpora; homepages; cyberpsychology; network analysis

1. Summary

The early Internet (e.g., content created and posted online during the mid-late 1990s) has received a fair amount of attention within the social-scientific literature. In particular, scholars [1,2] have examined the structure and content of newsgroups (text-based online communities divided by interest; e.g., alt.gossip.celebrities—a group devoted to gossip about celebrities). Much has been learned from the study of these groups.

However, as the Internet matured, different venues for personal expression emerged. One such venue was personal homepages: hypertext documents created by everyday users who wished to create a virtual space for themselves. Unlike newsgroups, which were interactive, full-duplex forums of interpersonal communication, personal homepages were unidirectional means of communication: they were created for broadcast distribution to a large audience with little or no means of providing feedback. To date, this type of personal expression on the early Internet has yet to be studied in detail.

The datasets described in the present paper is an attempt to ameliorate the lack of data on personal expression on the early Internet. Due to its popularity, I focus on the personal homepage hosting service Geocities. Founded in 1994, Geocities was one of the first widely adopted hosting providers to target individual Internet users [3]. The company operated independently from 1994 to 1999, at

which point it was acquired by Yahoo, Inc. This acquisition triggered concern about the long-term viability of Geocities webpages, fears which were confirmed in 2009, when Yahoo! announced its intention to shutter the service and remove all hosted websites [4]. In response, an independent group of digital archivists set out to create an archive of Geocities pages [5]. This archive was the basis for the present dataset.

2. Data Description

Geocities personal homepages can be divided into two broad categories: those created prior 1999 (when Yahoo! acquired the company) and after. Initially, Geocities followed a "neighborhood" structure, with each homepage residing within a larger neighborhood community. Pages were intended to fall within various themes: for example, "Area51" was a community for fans of science fiction, "WestHollywood" was a community for LGBT individuals, "CapitolHill" was designated for homepages related to government and politics. Each personal homepage was assigned an independent four-digit number. By then end of 1998, there were 42 separate neighborhoods, many of which had multiple "suburbs" (subcategories) of further-defined pages. Moreover, each neighborhood had its own community leadership—volunteers who assisted new users and helped to grow their respective neighborhoods [6]. This structure focused on community and proved to be a unique feature of Geocities.

When Yahoo! acquired Geocities in 1999, it instituted a new policy of "vanity" homepages, which followed a much simpler URL structure ("geocities.com/username"), abandoning the neighborhood structure [6]. As such, each personal homepage is identified either as a member of a neighborhood or an independent Yahoo! username. The present datasets contain data derived from both pre- and post-acquisition homepages.

The data described herein are derived from the Internet Archive's Geocities collection [7]. The complete archive is very large (approaching 1TB uncompressed), and contains many files not amenable to quantitative analysis (e.g., images, music files, etc.); as such, the described dataset is a reduced version of the complete archive that contains only hypertext (HTML) and ASCII text files, as well as derivatives of these data. This data deposit is divided into four parts, each of which can be retrieved separately (see https://osf.io/8u48d/):

1. a corpus of the raw text extracted from each homepage;
2. a psycho-linguistic analysis (using an open-source implementation [8] of the Linguistic Inquiry Word Count [9]) of the content of each personal homepage;
3. a dataset of nodes (Geocities websites) and (directed) edges (hyperlinks between Geocities websites), suitable for network analysis;
4. a manifest listing each file included in the corpus along with the date it was last modified.

All datasets can be merged with one another by using the homepage ID (see below) as a key. Details of each dataset are provided below. All files are compressed using the 7zip algorithm [10]. A depiction of the data extraction process is provided in Figure 1.

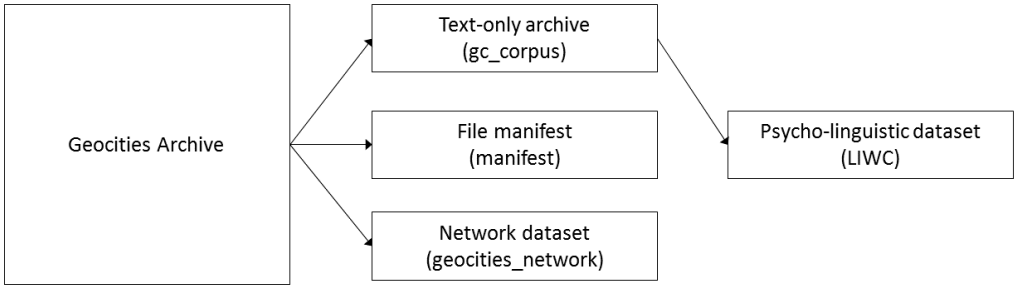

Figure 1. A visual depiction of the process used to create the datasets described in the present paper.

2.1. Raw Text Corpus

A raw text corpus is available [11], containing only the text of each homepage. This archive is unusually large (over 8 GB); as such, it is made available as a 7zip archive split into nine parts. To access the archive, all nine files will need to be downloaded and extracted. The uncompressed size of the corpus is approximately 42.2 GB.

The 7zip archive of the corpus contains two TAR archive files (each of which can be extracted using 7zip), "gc_neighborhood.tar" (pre-Yahoo! homepages) and "gc_yahooids.tar" (post-Yahoo! homepages). The neighborhood archive is structured in a straightforward manner: each neighborhood and suburb contains a text file of the text extracted from the combined pages for each homepage within that neighborhood. The Yahoo! ID archive is slightly more complex: because Windows systems are case-insensitive, Yahoo! IDs are divided into usernames that begin with uppercase and lowercase directories (so that they can be extracted and accessed on Windows machines), as well as a separate directory for usernames that begin with numbers.

Since the Geocities service was offered globally (indeed, the service is still active in Japan [12]), users from a variety of nationalities created content on the platform. As a result, valuable insights may be obtained by examining the language of individual homepages. This may be accomplished by applying a language detection library [13,14] to the raw text provided in this dataset. These data could, for example, be used as a proxy for the country in which a homepage was created.

2.2. Psycho-Linguistic Data

To facilitate easy analysis of the textual content of each homepage, a dataset of word frequency counts in 63 psycho-linguistic categories is provided [15]. This dataset was created using psyLex—an open-source implementation of the Linguistic Inquiry Word Count (LIWC) [8]. The LIWC is a simple technique used to extract the psychological properties of an author by counting the frequency of words in each category (e.g., positive emotion, negative emotion, aggression; a full description of each category and the process of their development is available from Pennebaker, Chung, Ireland, Gonzales, and Booth [16]).

This dataset contains an indicator of the relative URL ("file"), total word count ("wc"), and the percentage of words in the document that were classified ("percentClassified"). Subsequent fields represent raw word counts for each LIWC category.

The LIWC library employed was limited to English words, but other libraries are available from Pennebaker Conglomerates, Inc. [17] and may be applied to the raw text dataset (see Section 2.1). Within the psycho-linguistic dataset, non-English homepages were ignored.

2.3. Network Dataset

The network dataset is available [18] as a directed Graph Modeling Language (GML) file, compatible with most modern graph analysis packages and libraries. It contains a total of 622,482 separate nodes (individual homepages) with 756,039 edges (links with other Geocities homepages). The average degree (i.e., the number of links to and from an individual homepage) is 2.43.

Figure 2 shows a simplified network structure based on this dataset. The outer ring is comprised primarily of Geocities pages within the classic neighborhood system (discussed previously), while the inner core is comprised primarily of pages hosted under Yahoo!'s username system. A number of heavily weighted edges are visible between the inner and outer portions of the graph; these mostly represent users who combined accounts from both systems to create communities (multiple homepages) that were larger than would have been allowed by Geocities/Yahoo!

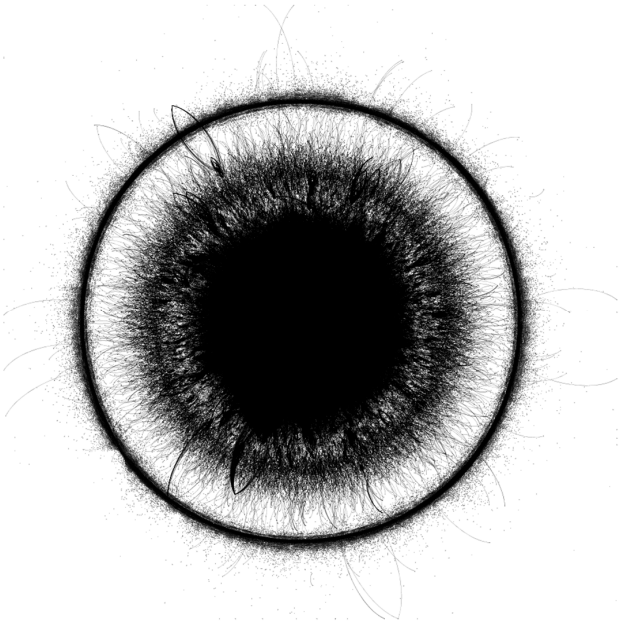

Figure 2. A visual representation of the links between Geocities websites using a ForceAtlas2 layout algorithm [19].

2.4. Manifest

The manifest [20] contains aggregate data on every file in the archive. It contains six fields, each of which is described in Table 1.

Table 1. Manifest fields.

Field Name	Description
Homepage	Homepage URL base [1]
Size	File size in bytes
Month	Month of last modification
Day	Day of last modification
Year	Year of last modification
File	File name

[1] Includes old (pre-Yahoo) Geocities neighborhood.

A total of 8,433,877 text or HTML files are included in the archive, with file sizes ranging from zero to 5.46 MB. Date stamps range from 1995 to 2009.

The manifest provides a means of assessing the amount of data provided by each homepage, as well as the most recent date on which a given page was modified. These data can be merged with any of the other files in the data deposit using a concatenation of the "Homepage" field and the "File" field and using the result as a unique identifier.

3. Methods

The data in this deposit were acquired by a set of independent archivists (the "Archive Team" [21]). Independent volunteers ran a set of scripts that crawled the Geocities website and downloaded the contents of each personal homepage. The URLs for targeted homepages were identified through two primary mechanisms: first, by systematically executing requests on Geocities neighborhood URLs, and second, by executing Google search queries on the geocities.com domain using a list of over 161,000 keywords (a collection of the scripts used to retrieve each homepage is available as part of this data deposit [22]). It is unknown exactly what percentage of homepages were retrieved [23], but the

final archive was approximately 900 GB. These data were then processed and made available as both a torrent [24] and through the Internet Archive [7]. Other websites [25] have made these data available in an accessible format.

This archive was retrieved over the course of nearly one year in 2012 and 2013. In order to verify that all files had been correctly retrieved, checksums were calculated and compared with those provided by the Internet Archive (www.archive.org) Geocities collection. Any corrupted or missing files were replaced with those retrieved directly from the Internet Archive. After all archives had been retrieved and validated, each was extracted using a series of linux shell scrips [11]. All non-text, non-HTML files were deleted. Each dataset described in the present paper was derived from this subset of the original Geocities archive.

To create the corpus and psycholinguistic dataset, it was necessary to remove HTML tags from all files with a .htm, .html, .HTM or .HTML extension. This was accomplished by flushing all HTML files through the w3m text browser and writing the output to a separate text file.

4. Potential Applications

The datasets described in the present paper can be employed in two broad ways: First, network and textual analyses can be used to form a better understanding of how human beings relate to one another and express themselves. The combination of network and psycho-linguistic data is a powerful starting place for social scientists. For example, in the context of a network analysis, Granovetter [26] has proposed that overlap between otherwise distal social networks can result in important social bonds. This hypothesis could be tested by evaluating the psycho-linguistic similarity of homepages that are proximally versus distally connected.

Second, the present dataset provides insight regarding personal expression and relationships on the Internet in its early years. The community-based structure of Geocities prior to Yahoo!'s acquisition of the service lends itself to network analysis, where community clusters may be clearly identified. For example, the linguistic properties of the "HollywoodHills" community might provide extensive insight into the online social support networks of LGBT individuals. Such results would coincide with previous studies of the use of electronic media as a support system for sexual minorities [28,29].

Alternately, extensive network analysis of the links between homepages might reveal differences between the types of communities that existed before and after Yahoo! did away with Geocities' formal neighborhood structure. Such an analysis could provide insight into formally imposed online community structures versus those which form naturally.

An even more informative analysis could be possible by combining the datasets provided: for example, one could examine the extent to which different modalities of emotional expression changed across time, and the extent to which diverse types of expression form distinct communities that can be identified through social network analysis.

In fact, the present data have already been used to further our understanding of how early online expression compared to more contemporary forms of electronic communication: the linguistic properties of Geocities personal homepages were analyzed and compared to those of Facebook and Twitter posts. This revealed that personal expression on the early Internet was, in fact, more intrinsically social (at least as measured by word counts) than more contemporary "social" network services [19,20]. This type of analysis could easily be extended and expanded by including network properties and longitudinal variables to further expand our understanding of how personal expression over the Internet has changed over the past 20 years.

It is my hope that other researchers will find ways to utilize this dataset that would be impossible to identify by a single researcher or lab.

Acknowledgments: The archive from which the current datasets were constructed was assembled by a group of independent archivists [21] and distributed with the help of the Internet Archive, and Jason Scott [30] in particular. I thank both groups for their efforts to compile and distribute this valuable dataset.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Adamic, L.A.; Buyukkokten, O.; Adar, E. A social network caught in the web. *First Monday* **2003**, *8*, 6. Available online: http://firstmonday.org/article/view/1057/977 (accessed on 19 December 2016). [CrossRef]
2. Hoffman, S. Processing Internet-derived Text—Creating a Corpus of Usenet Messages. *Lit. Linguist. Comput.* **2007**, *22*, 151–165. [CrossRef]
3. Available online: http://articles.latimes.com/1999/jan/29/business/fi-2730 (accessed on 30 December 2016).
4. Available online: http://arstechnica.com/web/news/2009/04/geocities-to-close-after-15-years-of-aesthetic-awesomeness.ars (accessed on 19 December 2016).
5. Available online: http://www.theregister.co.uk/2009/04/28/geocities_preservation/ (accessed on 19 December 2016).
6. Available online: http://www.bladesplace.id.au/geocities-neighborhoods-suburbs.html (accessed on 23 June 2013).
7. Available online: https://archive.org/web/geocities.php (accessed on 31 December 2016).
8. Available online: https://github.com/seanrife/psyLex (accessed on 24 November 2016).
9. Tausczik, Y.R.; Pennebaker, J.W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
10. Available online: http://www.7-zip.org (accessed on 19 December 2016).
11. Available online: https://osf.io/8u48d/files (accessed on 31 December 2016).
12. Available online: https://geocities.yahoo.co.jp/ (accessed on 30 December 2016).
13. Available online: https://pypi.python.org/pypi/langdetect (accessed on 30 December 2016).
14. Available online: https://cran.r-project.org/web/packages/textcat (accessed on 30 December 2016).
15. Available online: https://osf.io/535wz (accessed on 31 December 2016).
16. Available online: http://www.liwc.net/LIWC2007LanguageManual.pdf (accessed on 31 December 2016).
17. Available online: http://www.liwc.net (accessed on 30 December 2016).
18. Available online: https://osf.io/x5pwq (accessed on 31 December 2016).
19. Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **2014**, *9*, e98679. [CrossRef] [PubMed]
20. Available online: https://osf.io/g6xn6 (accessed on 31 December 2016).
21. Available online: http://www.archiveteam.org (accessed on 19 December 2016).
22. Available online: https://osf.io/khnqd/ (accessed on 31 December 2016).
23. Available online: http://ascii.textfiles.com/archives/2298 (accessed on 23 May 2017).
24. Available online: http://academictorrents.com/details/2dc18f47afee0307e138dab3015ee7e5154766f6 (accessed on 23 May 2017).
25. Available online: http://www.reocities.com/ (accessed on 23 May 2017).
26. Granovetter, M.S. The Strength of Weak Ties. *Am. J. Soc.* **1973**, *78*, 1360–1380. [CrossRef]
27. Rife, S.C. Personal expression in electronic media: A linguistic analysis of home pages and social media. In Proceedings of the Poster Session Presented at the Annual Meeting of the Southeastern Psychological Association, Hilton Head, SC, USA, 18–21 March 2015.
28. Haag, A.M.; Chang, F.K. The Impact of Electronic Networking on the Lesbian and Gay Community. *J. Gay Lesbian Soc. Serv.* **1997**, *7*, 83–94. [CrossRef]
29. Nieto, D.S. Who Is the Male Homosexual? A Computer-Mediated Exploratory Study of Gay Male Bulletin Board System (BBS) Users in New York City. *J. Homosex.* **1996**, *4*, 97–124. [CrossRef] [PubMed]
30. Available online: http://ascii.textfiles.com (accessed on 23 May 2017).