



# A Dataset of Late 1990s and Early 2000s Web Banner Ads on Chinese- and English-language Web Pages

**RICHARD LEWEI HUANG** 

**YUFENG ZHAO** 

\*Author affiliations can be found in the back matter of this article

DATA PAPER

**]**u[ubiquity press

## ABSTRACT

This dataset contains information about 22,915 banner advertisement images appearing on Chinese- and English-language web pages in the late 1990s and early 2000s archived on the Wayback Machine. For each ad image, the dataset provides information about the image's format and size, archived URLs of the image file, archived web pages the image appeared in, and, if available, web pages the image was linked to. The dataset also provides text data obtained from the images using optical character recognition (OCR). This dataset is useful for researchers in visual culture, history, media studies, and business and marketing.

## CORRESPONDING AUTHOR:

**Richard Lewei Huang**

The Information School,  
University of Washington,  
Seattle, WA, USA

[lw Huang@uw.edu](mailto:lw Huang@uw.edu)

---

## KEYWORDS:

web history; web archiving;  
online ads; online marketing;  
data art

## TO CITE THIS ARTICLE:

Huang, R. L., & Zhao, Y. (2024).  
A Dataset of Late 1990s and  
Early 2000s Web Banner  
Ads on Chinese-and English-  
language Web Pages. *Journal  
of Open Humanities Data*, 10:  
3, pp. 1–8. DOI: [https://doi.  
org/10.5334/johd.164](https://doi.org/10.5334/johd.164)

## (1) OVERVIEW

### REPOSITORY LOCATION

<https://doi.org/10.5281/zenodo.8408539>

### CONTEXT

This dataset contains information about 22,915 unique banner advertisements collected from Chinese- and English-language web pages from 1999 to 2003. Banner ads are a form of graphical advertisement prevalent on the World Wide Web in the late 1990s and early 2000s. The first banner ads appeared in 1994 on HotWired, a web-based digital publication venture launched by the Wired Magazine (McCullough, 2014). Imitating ads on magazines and billboards, banner ads are typically displayed across the top or sides of web pages and serve as clickable links that direct users to another web page serving advertising content.

Online banner ads have been a subject of scholarly inquiry in a variety of fields since their inception almost three decades ago. While the vast majority of existing scholarly literature on banner ads focuses on user interaction with the ads, especially factors that may influence user engagement with the ads (Burke et al., 2005; Lohtia et al., 2003; Resnick & Albert, 2014), there is also a growing body of literature that examines the cultures and histories of banner advertising (Ankerson, 2018; Jessen, 2010; Li & Zhunag, 2007). However, to date there has not been a systematic dataset of historical banner ad images openly available for researchers. While museums, archives, advertising firms, and independent archivists have long been collecting advertisements in different mediums, few conventional archives and collections have systematically documented or preserved web banner ads. In his 2018 review of 179 archives and collections of advertising, advertisement scholar Fred Beard found no established museum and university archives collecting digital advertisements. Among archives and collections maintained by advertisers, industry, and individual archivists, only nine of them included digital advertising in their collections (Beard, 2018). The only archive entirely dedicated to web advertising identified by Beard, Adverlicious,<sup>1</sup> is no longer accessible online.

In 2016, the Internet Archive launched GifCities, a search engine that allows the user to search for GIF images that exist in archived GeoCities web pages. A sizable amount of GIF files searchable on GifCities are banner ads, but GifCities only covers images appearing on GeoCities (jefferson, 2016). In 2018, independent archivist Tyler Grant released an archive of Flash-based banner ads that he manually downloaded from the Nielsen Ad Relevance database (Haskins, 2018). However, the archive does not contain any metadata about the downloaded Flash files, and the archive does not cover non-Flash banner ads. In addition, neither project has broad coverage of non-English banner ads.

In this paper, we present a dataset of web banner ads that grew out of a larger ongoing research project on Chinese-language web archiving. The aim of the original project is to measure and compare the archival rate and archival quality of Chinese- and English-language web pages from the late 1990s and early 2000s on the Wayback Machine. The project used printed Internet directory books published from that time period to collect historical URLs. Before the advent of full-text search engines, printed directory books of Internet resources were popular among web users to locate content of interest online (Ankerson, 2018). Formatted after phone books, the directory books usually provide lists of web page URLs manually curated into distinct categories. Today, the URLs featured in these books provide convenient entry points for researchers to access archived web content of the past. Using URLs found in six Internet directory books published in the United States and China in the late 1990s and early 2000s, we were able to access and download archived copies of web pages at these URLs on the Wayback Machine for the original research project. Since many archived web pages we downloaded contain banner ad images, we realised that it would be possible to build a dataset of historical banner ads on the web using the downloaded archived web pages and their metadata. We then devised a technical procedure to extract banner ad images from these archived web pages to compile this dataset.

---

<sup>1</sup> <http://adverlicio.us>.

## (2) METHOD

### STEPS

#### 1. Collecting banner ads

We collected 77,747 unique HTTP URLs from a collection of printed Internet directory books published in mainland China and the United States between June 1999 and December 2001 (See [Table 1](#)). The directory books published in mainland China feature mostly URLs of Chinese-language web pages in both simplified and traditional Chinese and a small number of English-language web pages, while the English-language web directory books feature mostly URLs in English.

| BOOK TITLE                                                                    | LANGUAGE           | ISBN          | PUBLISHER, PUBLICATION TIME, AND LOCATION                                          |
|-------------------------------------------------------------------------------|--------------------|---------------|------------------------------------------------------------------------------------|
| 全球中文网址速查手册/<br>Handbook of Chinese-<br>Language URLs *                        | Simplified Chinese | 9787542716002 | 上海科学普及出版社 / Shanghai<br>Popular Science Press *<br><br>1999/06<br><br>China        |
| Harley Hahn's Internet and<br>Web Yellow Pages, Millennium<br>Edition         | English            | 9780072121704 | McGraw-Hill Osborne Media<br><br>1999/09<br><br>United States                      |
| Que's Official Internet Yellow<br>Pages 2001 Edition                          | English            | 9780789724359 | Que Publishing<br><br>2000/09<br><br>United States                                 |
| 2001中文因特网黄页网址簿 /<br>2001 Chinese Internet Yellow<br>Pages *                   | Simplified Chinese | 9787900322241 | 上海电子出版有限公司 / Shanghai<br>Electronic Publishing Press *<br><br>2000/12<br><br>China |
| 全球中文网址速查手册（第<br>二版）<br><br>/ Handbook of Chinese-<br>Language URLs (2nd Ed) * | Simplified Chinese | 9787542718969 | 上海科学普及出版社 / Shanghai<br>Popular Science Press *<br><br>2001/01<br><br>China        |
| 2002中文因特网黄页网址簿 /<br>2002 Chinese Internet Yellow<br>Pages *                   | Simplified Chinese | 9787900348098 | 上海电子出版有限公司 / Shanghai<br>Electronic Publishing Press *<br><br>2001/12<br><br>China |

**Table 1** Internet directory books used to compile this dataset.

\*: English titles of Chinese-language books and names of their publishers are translated and provided by the authors here for reference only.

We used Wayback Machine's CDX API (<https://github.com/internetarchive/wayback/blob/master/wayback-cdx-server/README.md>) to get a list of available archived snapshots on the Wayback Machine of each URL in our list of URLs. Next, we used a Selenium script to fetch all web page snapshots made before January 1, 2003 and had an HTTP response status code of 200 (indicating that the snapshot was successfully captured by the Wayback Machine). We downloaded these snapshots in MHTML format using the web browser Chromium. MHTML is a web archive format that saves the HTML source code of a web page along with its embedded resources into one single file. Image files on a web page are commonly stored as base64-encoded strings in an MHTML file (Hopmann et al., 1999). If one URL had more than 50 snapshots with an HTTP 200 status code, we randomly sampled 50 snapshots to download. In total, 1,384,355 MHTML files were downloaded from the Wayback Machine.

We then used a Python script to analyse each MHTML file to extract image files (in GIF, JPEG, PNG, and BMP formats) with dimensions commonly used in banner ad images. The dimensions we used in our script are from the voluntary guidelines for Internet advertisements first

released in December 1996 by the Internet Advertising Bureau (IAB) (Collins, 1996). The IAB – whose name is now Interactive Advertising Bureau – is a trade association in digital advertising whose members include major brands, websites, ad agencies, and technology providers. The IAB Internet ad size guidelines were produced as an attempt to standardise Internet ad sizes, and they played an influential role in shaping the Internet advertising landscape through standardisation, which accelerated the growth of the ad network industry (Lobato & Thomas, 2020). As we wanted to focus on conventional horizontal banner ads, we opted to use specific IAB dimensions in our image extraction process (IAB, 1996, 2003; see Table 2).

|                |                                 |
|----------------|---------------------------------|
| 468 px * 60 px | Full Banner                     |
| 392 px * 72 px | Full Banner with Navigation Bar |
| 234 px * 60 px | Half Banner                     |
| 728 px * 90 px | Leaderboard                     |

**Table 2** Selected IAB banner sizes.

We calculated each image file's MD5 hash to group duplicate images that appear in multiple downloaded web page snapshots. For each unique image, we logged its appearances across our entire downloaded dataset of MHTML files, and for each appearance, we logged the image's archived URL in the Wayback Machine, the original URL of the web page containing the image, the timestamp of the archived web page snapshot, and, if available, the archived URLs to which the image is linked. In total, 22,915 unique images were extracted from all downloaded MHTML files.

## 2. OCR process

To improve reusability of the dataset, we used optical character recognition (OCR) software PaddleOCR (<https://github.com/PaddlePaddle/PaddleOCR>) to extract text in all banner ad images. PaddleOCR is an open-source OCR engine developed by the Chinese search engine company Baidu. In our experience, PaddleOCR delivered more accurate results for Chinese characters over other open-source OCR software, such as Tesseract. The OCR model version used by PaddleOCR in our text recognition process is PP-OCRv4. For animated GIF banner ad images, we separated the image into individual static frames, and extracted text from each frame of the image. In addition to text, we included in our dataset the confidence level and bounding box data provided by the OCR engine. We did not manually check the extracted text data for accuracy, because, in our experience, the existing data is already of reasonably high quality, and researchers can easily search for banners containing certain words using the OCR data.

## QUALITY CONTROL

In the OCR process, we performed file integrity checks on all GIF ad images and marked corrupted images in the dataset. The Wayback Machine may archive a corrupted image either because some kind of network error occurred when the Wayback Machine was trying to retrieve the image from the original server, or because the image was already corrupted on the original server, and the Wayback Machine was archiving the corrupted image verbatim. Corrupted images were detected in the process of separating animated images into frames using the image processing software library *imagemagick*. If *imagemagick* failed to separate an image into individual frames and reported it as corrupt, we would mark the image as corrupt in the dataset. No checks on file integrity on non-GIF images were performed, because PaddleOCR was able to process and output OCR results from all non-GIF images.

## DATA STRUCTURE

The dataset is presented as a JavaScript Object Notation (JSON) file containing an array of individual banner ad images. Each object in the array represents one unique banner ad image. Each object contains the following fields:

- **md5**: This field contains the MD5 hash value of the banner image file. It is used as a unique identifier for all banner ads in the dataset.
- **width** and **height**: These fields specify the dimensions of the banner ad in pixels.
- **filetype**: This field indicates the file format of the banner ad image as it was served from the original website (or the original ad provider's server) to the Wayback Machine's crawler. Possible values are `gif`, `jpeg`, `bmp`, and `png`. File type is detected by examining the first two characters of the image's base64 string.
- **appearances**: This is an array of objects. Each object represents one appearance of the banner ad in the downloaded collection of MHTML files along with associated details. An appearance is defined as the banner ad image located at a unique `image_url` (see below) appearing in a web page snapshot at a specific `url` archived at a specific `timestamp` (see below). Each object in this array contains the following fields:
  - **url**: This field provides the original URL of the web page where the banner ad was found.
  - **timestamp**: The timestamp indicates when the web page containing the banner ad was archived on the Wayback Machine. The timestamps are in the format of "YYYYMMDDHHMMSS". The archived snapshot of the web page containing the banner ad image can be accessed at <https://web.archive.org/web/{{timestamp}}/{{url}}>
  - **image\_url**: This field provides the archived URL to the banner ad image as it appeared in the archived snapshot of the web page captured at the time indicated in `timestamp`.
  - **hrefs**: This field is an array containing archived URLs that the image would lead the user to upon clicking. In most cases, the array contains only one element. If this array contains multiple elements, it indicates that the banner image loaded from the same `image_url` appeared on the archived snapshot of the web page multiple times and was linked to at least two different URLs.
- **ocr\_result**: If the image is not corrupted, `ocr_result` is an array containing text extracted from the image using PaddleOCR. For animated images, each object in this array represents one individual frame. For static images, there is only one object in this array, with `frame_num` (see below) being 0. For corrupted images, the value of `ocr_result` will be "corrupt". If the image is not corrupted, an object in this array contains the following fields:
  - **frame\_num**: The number of the specific frame of the banner ad image that this object is representing (counting from zero).
  - **result**: an array representing bounding boxes detected by the OCR engine on the frame. A bounding box is a rectangular area in the image containing text as detected by the OCR engine. Each object in this array contains the following fields:
    - **text**: the text detected by the OCR engine.
    - **confidence**: the confidence score given by the OCR engine for the text detected. The value is between 0 and 1, with a higher score indicating a higher level of certainty the OCR engine has regarding the accuracy of the OCR results.
    - **bounding\_box**: an array containing 4 sub-arrays representing the coordinates of the four corners of the bounding box.

### (3) DATASET DESCRIPTION

**Object name** banners\_output\_20230930.json

**Format names and versions** JavaScript Object Notation (JSON)

**Creation dates** 2023-09-30

**Dataset creators** Richard Lewei Huang, Yufeng Zhao

**Language** All variable names are in English. Most banner ads are in either English or Chinese (simplified and traditional).

## (4) REUSE POTENTIAL

We expect this dataset to be useful for researchers from a variety of disciplines who are interested in analysing banner ad images using different methods. By studying the evolution of banner ads in this dataset, researchers can gain insights into the changing aesthetics of online advertisements and the ways in which banner ads reflected the broader socio-cultural context of the web at the time. Since the banner ads are collected from Chinese- and English-language web pages, comparative studies of banner advertising along linguistic and cultural differences are also possible through analysis of the dataset.

Additionally, the dataset may be of interest to artists looking to create data-based artwork using the banner ads. As an example, we have created Banner Depot 2000, a website<sup>2</sup> where visitors can browse through the banner ads dataset, search for specific banners by keyword, and compose “found poetry” using individual frames of banner ad images in the dataset as poetry verses. Banners on the website are displayed on a fair use basis for the purpose of scholarly research and criticism, as well as transformative creative expression.

## LIMITATIONS AND FUTURE WORK

We did not include the actual banner ad image files in our dataset due to copyright concerns. However, interested researchers should be able to download any image file manually using the value of the field `image_url` in any object under the `appearances` array in an ad image object in the dataset.

Given the linguistic and thematic diversity of the banner ads, we decided not to provide any kind of subjective metadata for the ads. With the OCR data as well as other types of data provided in the dataset, interested researchers should be able to categorise and classify the banner ads in different ways that make sense for their own research projects. We are also considering building a tagging and collections feature into our website (see footnote 2) where users can add tags to banners and share customised collections of banner ads, thus enabling collective annotation and filtering of the dataset.

We are not providing manually corrected OCR data in this version of the dataset, because, in our experience, the quality of the existing OCR data is already reasonably high to support keyword searching within the dataset, and performing additional manual error correction may provide marginal benefits that would not justify the potential amount of human labour required for the task. In future versions of the dataset, we will consider providing multiple versions of OCR data generated by different OCR engines in lieu of manually corrected data to help researchers who need more reliable text transcriptions of the banner ads.

Flash was a commonly used technology for displaying banner ads in the late 1990s and early 2000s. However, our dataset currently does not contain any Flash-based banner ads. This is because Chromium does not include Flash files in a web page when saving it to an MHTML file. We will consider developing a new mechanism to detect Flash ads in our collection of archived web pages and incorporate them into a future version of the dataset.

Due to technical limitations in our data curation methods, we are also unable to provide any metrics commonly used in the online advertising industry, such as impressions, click-through rates, and cost per click.

Since many banner ads are served by ad networks, which determine what ads are displayed for a user based on a variety of factors (such as the user’s geographic location, the content of the web page, and the browser the user is using), the specific hardware, software, and network configurations of the Wayback Machine’s crawler might have inadvertently played a role in

---

2 <https://www.banner-depot-2000.net>.

determining what ads were archived by the Wayback Machine. Researchers must keep this in mind as they interpret any potential findings through analysis of the dataset.

## ACKNOWLEDGEMENTS

The authors want to thank University of Washington undergraduate students Yunxuan Wu and Tianhao Yao for their assistance in data curation. The authors also want to thank Nicholas Weber, Eva Maxfield Brown, Jack B Du, and Han Su for their suggestions and comments.

## FUNDING INFORMATION

This research was supported in part by a grant from New America's Public Interest Technology University Network program, and the University of Washington's Information School.

## COMPETING INTERESTS

The authors have no competing interests to declare.


## AUTHOR CONTRIBUTIONS

Richard Lewei Huang: conceptualization, data curation, methodology, writing-original draft, software.

Yufeng Zhao: data curation, methodology, software.

## AUTHOR AFFILIATIONS

**Richard Lewei Huang**  [orcid.org/0000-0002-0264-9300](https://orcid.org/0000-0002-0264-9300)  
The Information School, University of Washington, Seattle, WA, USA

**Yufeng Zhao**  [orcid.org/0009-0006-8479-5497](https://orcid.org/0009-0006-8479-5497)  
Independent researcher, Brooklyn, NY, USA

## REFERENCES

- Ankerson, M. S.** (2018). Cool Quality and the Commercial Web (1994–1995). In *Dot-Com Design: The Rise of a Usable, Social, Commercial Web*. New York: NYU Press.
- Beard, F.** (2018). Archiving the archives: The world's collections of historical advertisements and marketing ephemera. *Journal of Historical Research in Marketing*, 10(1), 86–106. DOI: <https://doi.org/10.1108/JHRM-08-2017-0044>
- Burke, M., Hornof, A., Nilsen, E., & Gorman, N.** (2005). High-cost banner blindness: Ads increase perceived workload, hinder visual search, and are forgotten. *ACM Transactions on Computer-Human Interaction*, 12(4), 423–445. DOI: <https://doi.org/10.1145/1121112.1121116>
- China Telecom & Shanghai Telephone Directory Corporation.** (2000). *2001 yin te wang huang ye wang zhi bu* [2001 Chinese Internet Yellow Pages]. Shang Hai Dian Zi Chu Ban You Xian Gong Si.
- China Telecom & Shanghai Telephone Directory Corporation.** (2001). *2002 yin te wang huang ye wang zhi bu* [2002 Chinese Internet Yellow Pages]. Shang Hai Dian Zi Chu Ban You Xian Gong Si.
- Collins, G.** (1996, December 12). Trade Groups Propose Web Banner Guidelines. *The New York Times*. Retrieved from <https://www.nytimes.com/1996/12/12/business/trade-groups-propose-web-banner-guidelines.html> (last accessed: 29 September 2023).
- Hahn, H.** (2000). *Harley Hahn's Internet & Web Yellow Pages*. New York: Osborne/McGraw-Hill.
- Haskins, C.** (2018, October 8). This 20-Year-Old Is Archiving Thousands of Flash Banner Ads From the Early 2000s. *Vice*. Retrieved from <https://www.vice.com/en/article/7x3d7d/flash-banner-ad-archive> (last accessed: 28 September 2023).
- Hopmann, A., Palme, J., & Shelness, N. H.** (1999). *MIME Encapsulation of Aggregate Documents, such as HTML (MHTML)* (Request for Comments RFC 2557). Internet Engineering Task Force. DOI: <https://doi.org/10.17487/RFC2557>
- IAB.** (1996, December). *IAB/CASIE PROPOSAL FOR VOLUNTARY MODEL BANNER SIZES* [Press release]. Retrieved from <https://web.archive.org/web/19980211033307/http://www.iab.net/news/content/library/decbanner.html> (last accessed: 29 September 2023).
- IAB.** (2003, April 8). *IAB Standards and Guidelines—Ad Unit Guidelines*. Retrieved from <http://web.archive.org/web/20030408033710/http://www.iab.net/standards/adunits.asp#> (last accessed: 29 September 2023).



- jefferson.** (2016, November 1). GifCities: The GeoCities Animated GIF Search Engine. *Internet Archive Blog*. Retrieved from <https://blog.archive.org/2016/11/01/gifcities-the-geocities-animated-gif-search-engine/> (last accessed: 28 September 2023).
- Jessen, I. B.** (2010). The Aesthetics of Web Advertising: Methodological Implications for the Study of Genre Development. In N. Brügger (Ed.), *Web History* (pp. 257–279). New York: Peter Lang.
- Li, X., & Zhunag, L.** (2007). Cultural Values in Internet Advertising: A Longitudinal Study of the Banner Ads of the Top U.S. Web Sites. *Southwestern Mass Communication Journal*, 23(1), 57–72.
- Lobato, R., & Thomas, J.** (2020). Formats and formalization in internet advertising. In M. Jancovic, A. Volmar, & A. Schneider (Eds.), *Format matters: Standards, practices, and politics in media cultures* (pp. 65–80). Lüneberg: meson press.
- Lohtia, R., Donthu, N., & Hershberger, E. K.** (2003). The Impact of Content and Design Elements on Banner Advertising Click-through Rates. *Journal of Advertising Research*, 43(4), 410–418. DOI: <https://doi.org/10.2501/JAR-43-4-410-418>
- McCullough, B.** (2014, October 27). On The 20th Anniversary, An Oral History of the Web's First Banner Ads. *Internet History Podcast*. Retrieved from <https://www.internethistorypodcast.com/banner/> (last accessed: 20 September 2023).
- Resnick, M., & Albert, W.** (2014). The Impact of Advertising Location and User Task on the Emergence of Banner Ad Blindness: An Eye-Tracking Study. *International Journal of Human-Computer Interaction*, 30(3), 206–219. DOI: <https://doi.org/10.1080/10447318.2013.847762>
- Shang Hai Da Xue Li Pu Wang Luo Jiao Yu Zhong Xin.** (1999). *Quan qiu zhong wen wang zhi su cha shou ce* [Handbook of Chinese-Language URLs]. Shang Hai Ke Xue Pu Ji Chu Ban She.
- Shang Hai Da Xue Li Pu Wang Luo Jiao Yu Zhong Xin.** (2001). *Quan qiu zhong wen wang zhi su cha shou ce (di er ban)* [Handbook of Chinese-Language URLs, Second Edition]. Shang Hai Ke Xue Pu Ji Chu Ban She.
- Turner, M. L., & Seybold, A.** (2000). *Que's Official Internet Yellow Pages*. Indianapolis: Que.

#### TO CITE THIS ARTICLE:

Huang, R. L., & Zhao, Y. (2024). A Dataset of Late 1990s and Early 2000s Web Banner Ads on Chinese-and English-language Web Pages. *Journal of Open Humanities Data*, 10: 3, pp. 1–8. DOI: <https://doi.org/10.5334/johd.164>

**Submitted:** 06 October 2023

**Accepted:** 27 November 2023

**Published:** 03 January 2024

#### COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.