# Digital collections: If you build them, will they visit?

**Frederick Zarndt**
**IFLA Newspapers Section, Coronado CA USA**
frederick@frederickzarndt.com

**Brian Geiger**
**California Digital Newspapers Collection, University of California Riverside, Riverside CA USA**
bgeiger@ucr.edu

**Robert Stauffer**
**Hoʻolaupaʻi Hawaiian Nūpepa Collection, Honolulu HI USA**
bob@rstauffer.com

**Alyssa Pacy**
**Cambridge Public Library, Cambridge MA USA**
apacy@cambridgema.gov

**Meredith Palmer**
**DL Consulting, Hamilton, New Zealand**
meredith@dlconsulting.com

**Joanna DiPasquale**
**Vassar College, Poughkeepsie NY USA**
jdipasquale@vassar.edu

## Abstract:

*How do your cultural heritage organization's digital collections fare in search rankings? Assuming your collections have newspapers from 1915, will a Google search for information about the "Battle of Gallipoli" return results? At the April 2012 Bibliothèque nationale de France International Newspapers Conference, one of the authors examined web traffic rankings and search results for digital newspaper collections at libraries around the world. Both traffic rankings and search results showed that content in cultural heritage organizations' digital collections dwell in Internet obscurity (http://bit.ly/parisinternationalnewspapers).*

*In this session we re-visit these rankings and results, examining what it means for a digital collection to be successful. Is success only about page views, unique visitors, and bounce rates? Paraphrasing Trevor Owens blog*[1]*, if the mission of a cultural heritage organization is more than random users flipping through the pages of its digital collections, how does one encourage and measure community engagement? Is crowdsourcing "the single greatest advancement in getting people using and interacting with library collections"?*

*We describe simple methods some organizations are using to market their collections and engage their users by combining a balanced mix of digital, social and print media and leveraging their primary marketing tool, the collection itself.*

**Keywords:** newspapers, digital historical newspaper collections, digital collections marketing, SEO, genealogy

## INTRODUCTION

Digital newspaper collection custodians **Brian Geiger** of the California Digital Newspapers Collection, University of California Riverside in California, **Robert Stauffer** of Hoʻolaupaʻi Hawaiian Nūpepa Collection in Hawaii, **Alyssa Pacy** of Cambridge Public Library in Cambridge Massachusetts, and **Joanna DiPasquale** of Vassar College in Poughkeepsie New York, share a challenge common to digital collection owners worldwide. How do you create visibility for and drive public awareness of a digital collection so that it is easy to discover and access? The authors of this paper tell their stories and provide insight into lessons learned that can be replicated for other digital collections whether they are in the early stages of planning, well into development, or fully established. **Frederick Zarndt**, Chair of the IFLA Newspaper Section, opens and frames the discussion by offering an interpretation of current search ranking and site traffic analytics for digital collections around the world. Frederick and his co-authors examine what it means for a digital collection to be successful.

## 1. WE BUILT IT. WHY IS NO ONE VISITING?

It's no longer news: Digital collections at libraries everywhere are multiplying and expanding at rates limited only by library budgets. Historical newspaper collections are especially in favor, because, as surrogates for the original newspapers, digital newspapers are far easier to use and reach a much broader and more dispersed user population. But do digitized historic newspapers live up to their potential?

At the 2012 International Newspaper Conference organized and hosted by Bibliothèque nationale de France, one of the authors explored the Internet search visibility of cultural heritage organization (CHO) digital newspaper collections (DNC) using "outside-in" web analytics (Alexa) and with a specific search for a major historical event (the Battle of Gallipoli) which would have been mentioned in any newspaper in the world when it occurred (1915-1916). The exploration showed that while historical DNCs are popular in some countries, in other countries they are (much) less popular. And in all countries the collections were virtually invisible to a Google search for information about the Battle of Gallipoli. Invisible even though DNCs are accurate facsimiles of the newspapers of that time, are

[1] Paraphrased from Trevor Owen's blog http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/ (accessed June 2013).

primary sources for other, secondary Internet articles, and have many, many stories about and references to the event.

Has Internet search visibility of these collections changed since the Conference? No, not at all. And perhaps, with the demise of Google News, their visibility has even declined. But more about this shortly.

First, let's ask who uses DNCs. Overwhelmingly the answer is genealogists and family historians who are 50+ years of age.

In order to learn about their DNC user demographic, the California Digital Newspaper Collection (CDNC) and the Cambridge Public Library surveyed users of their collections[2]. A similar survey was done by Utah Digital Newspapers with similar results[3]. Less formal and older surveys at the National Library of New Zealand and the National Library of Australia show a similar user demographic[4].

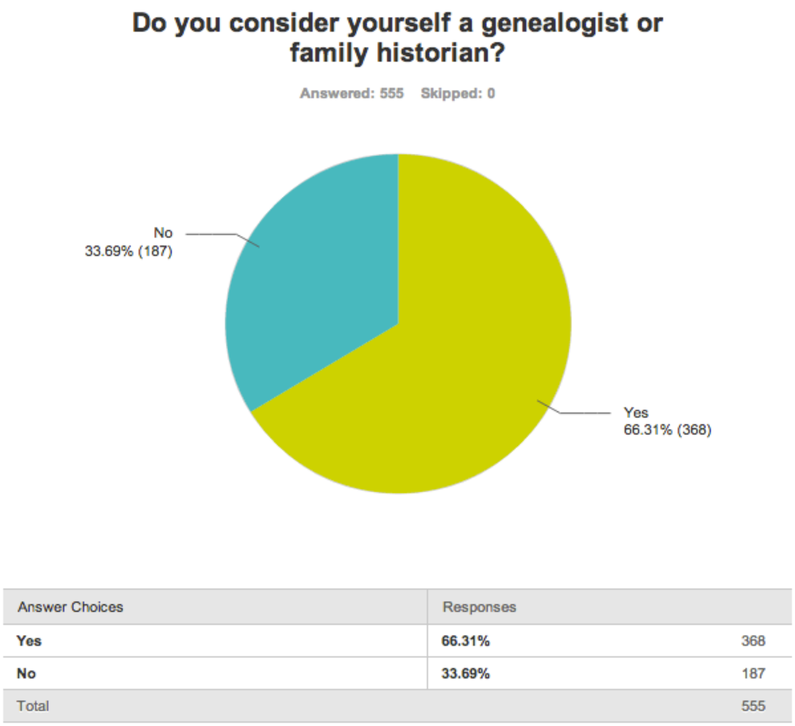
From February to May 2013 the CDNC user survey got 555 responses. From January to May 2013 the Cambridge survey got 30 responses. For both collections users are overwhelmingly genealogists / family historians: 82% for Cambridge and 66% for CDNC.

As for age, 75% of Cambridge users are 50+ years old; nearly 80% of CDNC users are 50+ years old. Survey results also show that Cambridge has no users under 30 years of age and fewer than 5% of CDNC users are under 30. The majority of Papers Past and Trove newspaper collection users are also 50+ years of age according to library surveys.

[2] The CDNC survey questions can be found at the end of this paper. The Cambridge questions are nearly identical except for the name of the collection.

[3] Randy Olsen and John Herbert. *Small town papers: still delivering the news*. World Library and Information Congress. Helsinki, Finland. August 2012. http://conference.ifla.org/past/ifla78/session-119.htm (accessed June 1, 2013).

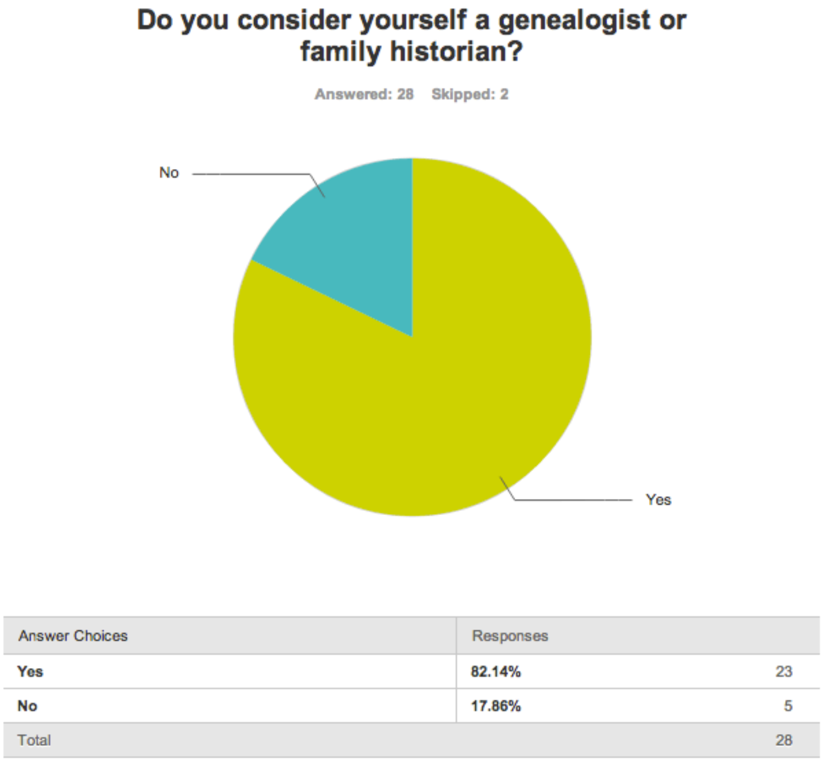
[4] At present both the National Library of New Zealand and National Library of Australia are surveying their users and may shortly update and revise their DNC user profile. National Library of Australia users are mentioned in the Library's 2012 Trove Annual Report (http://www.nla.gov.au/librariesaustralia/files/2011/05/08-Trove-Report-2012-updated.pdf).



### Do you consider yourself a genealogist or family historian?

Answered: 555 Skipped: 0

| Answer Choices | Responses | |
|---|---|---|
| **Yes** | **66.31%** | 368 |
| **No** | **33.69%** | 187 |
| Total | | 555 |



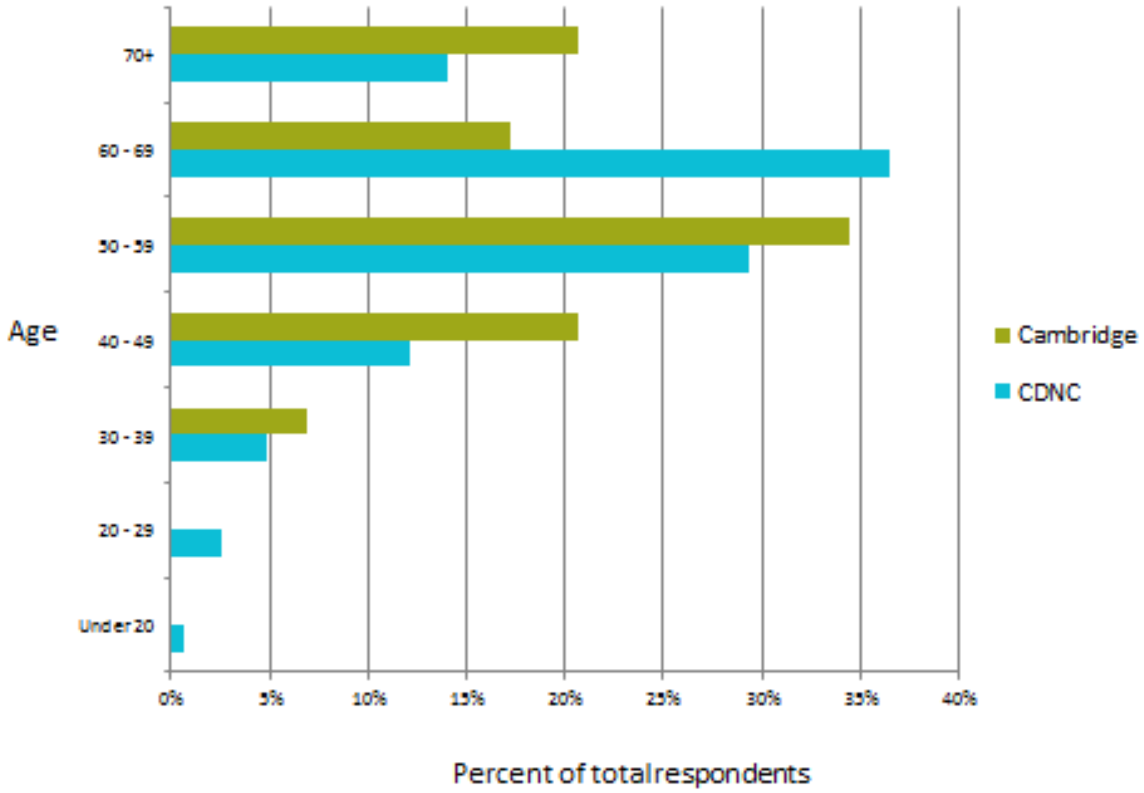
### Do you consider yourself a genealogist or family historian?

Answered: 28 Skipped: 2

| Answer Choices | Responses | |
|---|---|---|
| **Yes** | **82.14%** | 23 |
| **No** | **17.86%** | 5 |
| Total | | 28 |

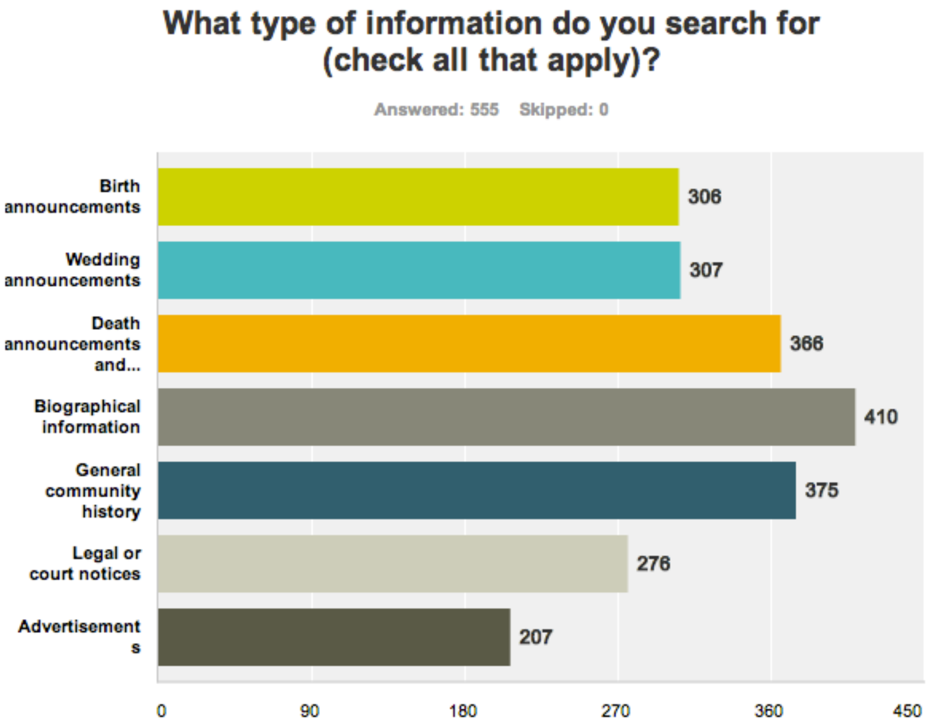
CDNC User Demographic

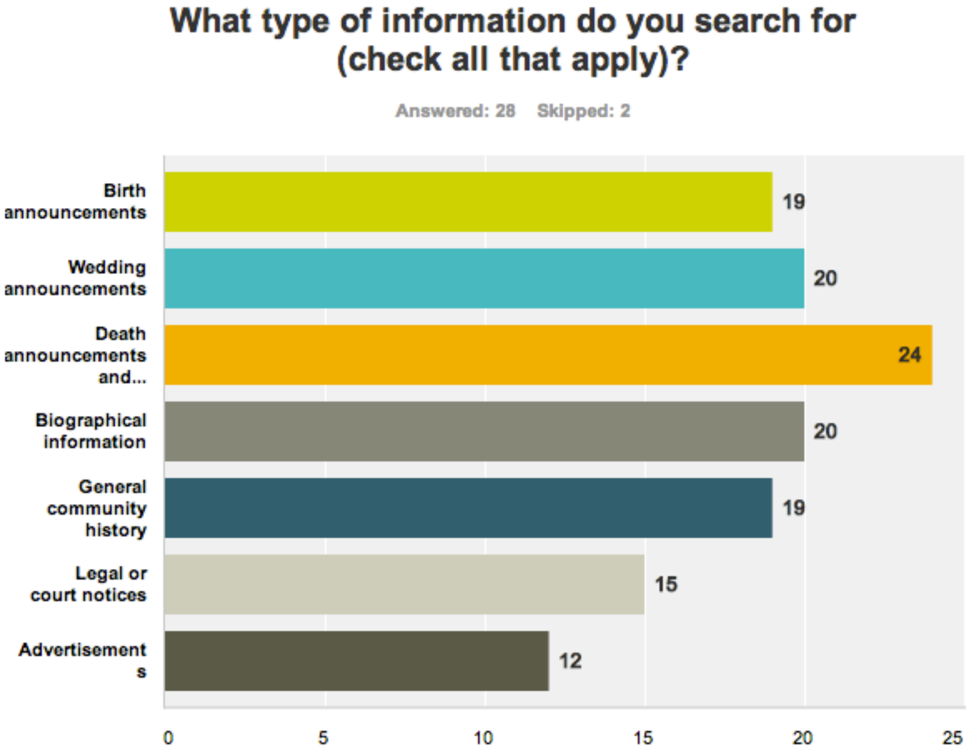
Cambridge User Demographic



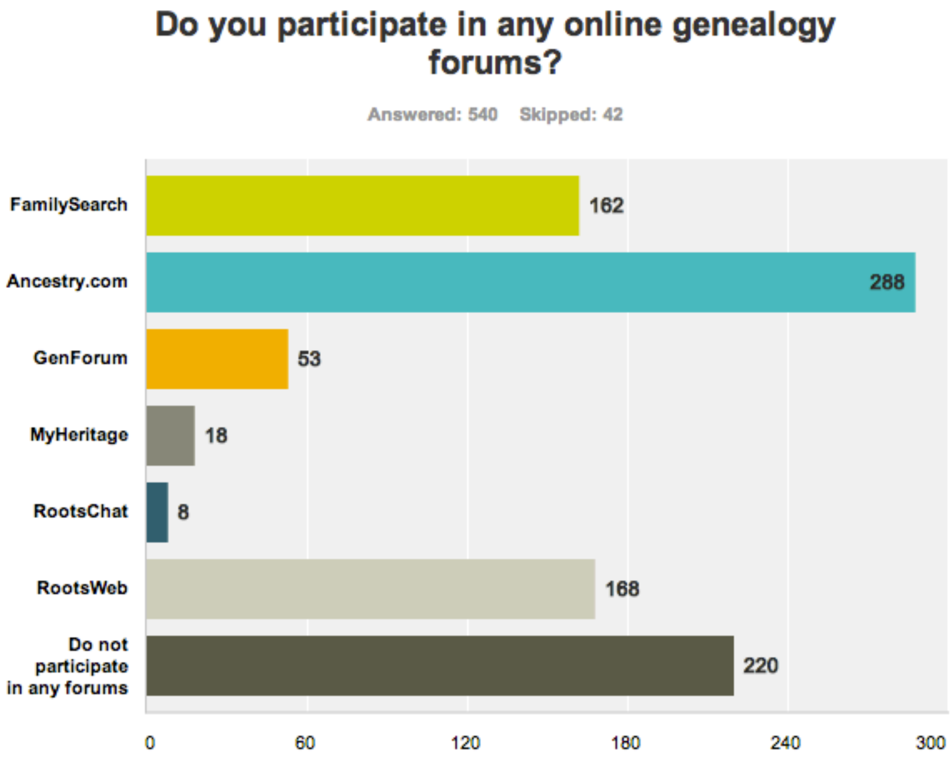
Both the CDNC and Cambridge surveys ask what sort of stuff is most interesting to users. The graphs below show what one would expect for users interested in genealogy: They search mostly for obituaries, general family announcements (births, wedding), and biographical information. Olsen and Herbert's Utah Digital Newspapers user surveys report similar results[5].

[5] Olsen and Herbert. *Small town papers: still delivering the news*.





The CDNC survey asks its users if they participate in or subscribe to a genealogy services like FamilySearch, Ancestry.com, RootsWeb, etc. Survey results show that about 60% use one or more one or more such services.



In their research it is unlikely that genealogists will use only these services (FamilySearch, Ancestry.com, RootsWeb,…) and DNC collections at CDNC, Cambridge, and other libraries. Where else would they search? Obviously Google (or Bing or Yahoo! or some other search engine). Although we will not replicate a search for specific family information in this paper, we can do a similar search to show the (in)visibility of historical digital newspaper collections to common search engines.

For the 2012 International Newspaper Conference, one of the authors searched for information about a historical event from 1915-1916, the Battle of Gallipoli. This particular event was chosen because newspapers around the world reported it and because these same newspapers are now all out-of-copyright[6]. The search used was:

(battle OR campaign) AND (Gallipoli OR Dardenelles OR Çanakkale)

with a date range 1-Jan-1915 to 31-Dec-1916[7].

The results? Not a single hit from a library historical digital news collection within the 1st 100 results!

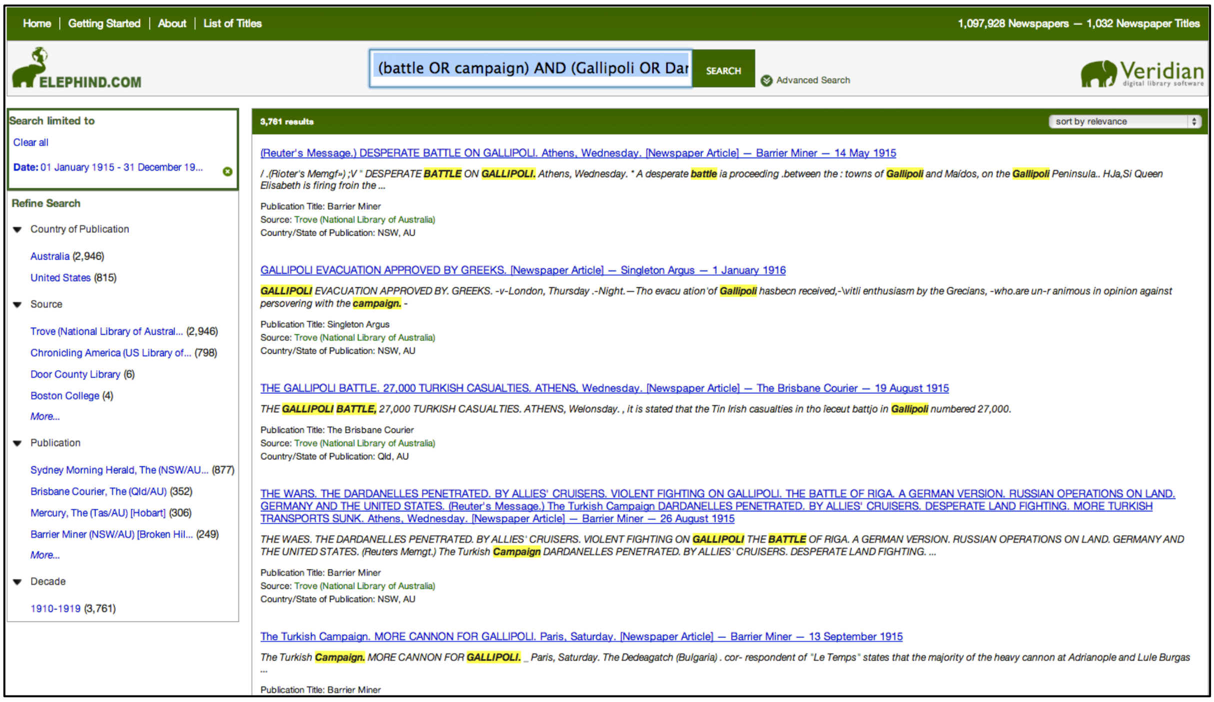
Thinking that a search focused on news articles alone would be more fruitful, the same search was repeated at various Google News sites: http://news.google.com,

[6] Out-of-copyright is an important consideration since DNCs consist mostly of out-of-copyright newspapers.

[7] Note that neither a simple Google search nor an advanced search any longer allows one to specify a date range as specifically as 1-Jan-1915 to 31-Dec-1916. However it is still possible with Google News (http://news.google.com) search.

http://news.google.com.au, http://news.google.com.sg, and http://news.google.co.nz. As with the simple search, again there was not a single hit from a library DNC in the 1st 100 results! Not unexpectedly most hits were from Google News itself, but there were also many hits were from other collections, both free and behind paywalls (New York Times, Atlanta Constitution, Boston Daily Globe). Note that although the 2012 search results through Google News had several results from the National Library of New Zealand's Papers Past newspapers collection, this year's searches yielded not a single hit from Papers Past or from any other CHO digital newspaper colleciton.

If one performs this same search at each digital newspaper collection website, and there is no dearth of results. Or, instead of searching each collection separately, simply use Elephind[8]. Elephind has not yet indexed all pages in every historical DNC, but it is clear from the screen shot of the search, that the collections have plenty of references to the Battle of Gallipoli.


Home | Getting Started | About | List of Titles
1,097,928 Newspapers — 1,032 Newspaper Titles
ELEPHIND.COM
(battle OR campaign) AND (Gallipoli OR Dar
SEARCH
Advanced Search
Veridian
digital library software

Search limited to

Clear all

Date: 01 January 1915 - 31 December 19...

Refine Search

- Country of Publication
  - Australia (2,946)
  - United States (815)
- Source
  - Trove (National Library of Austral... (2,946)
  - Chronicling America (US Library of... (798)
  - Door County Library (6)
  - Boston College (4)
  - More...
- Publication
  - Sydney Morning Herald, The (NSW/AU... (877)
  - Brisbane Courier, The (Qld/AU) (352)
  - Mercury, The (Tas/AU) [Hobart] (306)
  - Barrier Miner (NSW/AU) [Broken Hil... (249)
  - More...
- Decade
  - 1910-1919 (3,761)

3,761 results

sort by relevance

(Reuter's Message.) DESPERATE BATTLE ON GALLIPOLI. Athens, Wednesday. [Newspaper Article] — Barrier Miner — 14 May 1915

*/ .(Rioter's Memgf») ;V " DESPERATE **BATTLE** ON **GALLIPOLI.** Athens, Wednesday. " A desperate **battle** ia proceeding .between the : towns of **Gallipoli** and Maídos, on the **Gallipoli** Peninsula.. HJa,Si Queen Elisabeth is firing froin the ...*

Publication Title: Barrier Miner
Source: Trove (National Library of Australia)
Country/State of Publication: NSW, AU

GALLIPOLI EVACUATION APPROVED BY GREEKS. [Newspaper Article] — Singleton Argus — 1 January 1916

***GALLIPOLI** EVACUATION APPROVED BY. GREEKS. -v-London, Thursday .-Night.—Tho evacu ation'of **Gallipoli** hasbecn received,-\vitli enthusiasm by the Grecians, -who.are un-r animous in opinion against persovering with the **campaign.** -*

Publication Title: Singleton Argus
Source: Trove (National Library of Australia)
Country/State of Publication: NSW, AU

THE GALLIPOLI BATTLE. 27,000 TURKISH CASUALTIES. ATHENS, Wednesday. [Newspaper Article] — The Brisbane Courier — 19 August 1915

*THE **GALLIPOLI BATTLE,** 27,000 TURKISH CASUALTIES. ATHENS, Welonsday. , it is stated that the Tin Irish casualties in tho ïeceut battjo in **Gallipoli** numbered 27,000.*

Publication Title: The Brisbane Courier
Source: Trove (National Library of Australia)
Country/State of Publication: Qld, AU

THE WARS. THE DARDANELLES PENETRATED. BY ALLIES' CRUISERS. VIOLENT FIGHTING ON GALLIPOLI. THE BATTLE OF RIGA. A GERMAN VERSION. RUSSIAN OPERATIONS ON LAND. GERMANY AND THE UNITED STATES. (Reuter's Message.) The Turkish Campaign DARDANELLES PENETRATED. BY ALLIES' CRUISERS. DESPERATE LAND FIGHTING. MORE TURKISH TRANSPORTS SUNK. Athens, Wednesday. [Newspaper Article] — Barrier Miner — 26 August 1915

*THE WAES. THE DARDANELLES PENETRATED. BY ALLIES' CRUISERS. VIOLENT FIGHTING ON **GALLIPOLI** THE **BATTLE** OF RIGA. A GERMAN VERSION. RUSSIAN OPERATIONS ON LAND. GERMANY AND THE UNITED STATES. (Reuters Memgt.) The Turkish **Campaign** DARDANELLES PENETRATED. BY ALLIES' CRUISERS. DESPERATE LAND FIGHTING. ...*

Publication Title: Barrier Miner
Source: Trove (National Library of Australia)
Country/State of Publication: NSW, AU

The Turkish Campaign. MORE CANNON FOR GALLIPOLI. Paris, Saturday. [Newspaper Article] — Barrier Miner — 13 September 1915

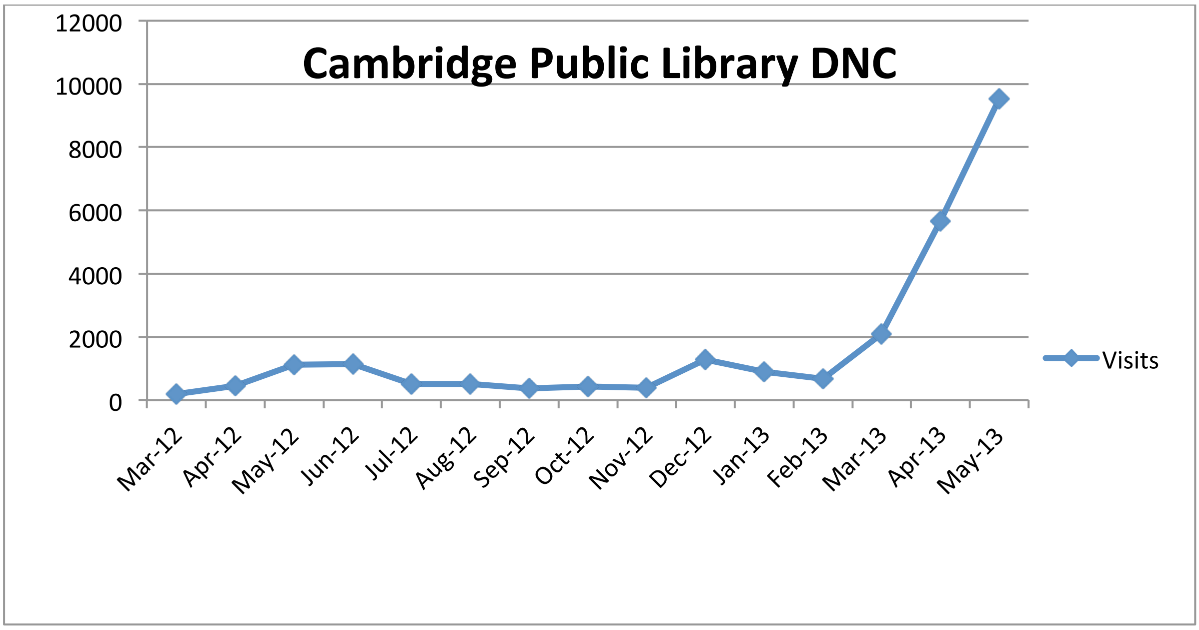
*The Turkish **Campaign.** MORE CANNON FOR **GALLIPOLI.** _ Paris, Saturday. The Dedeagatch (Bulgaria) . cor- respondent of "Le Temps" states that the majority of the heavy cannon at Adrianople and Lule Burgas ...*

Publication Title: Barrier Miner

A simple Google search is exactly how unskilled researchers will start and end their family history search on the Internet. Even experienced researchers, if they do not know about the DNC websites -- and some are well-hidden -- may never find the valuable information in documents considered to be "the first draft of history".

[8] As Elephind's about page says "..it is now possible to search digital newspaper collections from around the globe in the aggregate. Elephind.com is much like Google, Bing, or other search engines but focused on only historical, digitized newspapers. By clicking on the Elephind.com search result that interests you, you'll go directly to the newspaper collection which hosts that story". Visit its about page to see which collections are currently included. If your collection isn't among them, email the host (also from the about page) to ask for its inclusion.

What's the reason for such poor search results from a simple Google search for major historical event? To give a simple, but not complete, answer this question, let's look at Google analytics for the Cambridge Public Library's digital historical newspaper collection.



Prior to February 2013 the Cambridge Library DNC was in the same state as most: It was invisible to the big search engines. In February DL Consulting changed Cambridge's newspaper collection robots.txt file and added a XML sitemap file. These changes had, as one can see in the web analytics graph above, a dramatic effect on the number of visits from "organic" searches[9].

Is increasing DNC website traffic really this simple? Let's have a closer look at Cambridge's newspaper collection as well as those from Hawaii, California Digital Newspaper Collection, and Vassar College.

## 2. MARKETING HAWAIIAN NEWSPAPERS

The Hoʻolaupaʻi Hawaiian Nūpepa Collection is a digital newspaper archive located within Ulukau, the Hawaiian Electronic Library. The purpose of Ulukau is to make resources available for the use, teaching, and revitalization of the Hawaiian language and for a broader and deeper understanding of Hawaiʻi. The Ulukau library consists of a library of digital books (which come with special digital tools), and also a series of special collections including photographs, curriculum materials, genealogy records, music, dictionaries, newspaper archives, and much more. The site was founded by the Hale Kuamoʻo Center for Hawaiian Language and Culture Through the Medium of Hawaiian (a department within Ka Haka ʻUla O Keʻelikōlani College of Hawaiian Language at the University of Hawaiʻi at

[9] *Organic search* results are listings on search engine results pages that appear because of their relevance to the search terms as opposed to their being advertisements. Definition from Wikipedia (http://en.wikipedia.org/wiki/Organic_search) accessed June 2013.

Hilo). Ulukau is co-sponsored by Hale Kuamoʻo and the Native Hawaiian Library at ALU LIKE, Inc.

The Hawaiian language is one of the surviving indigenous languages within the boundaries of the United States. During the period 1834 until 1948, over 100 different Hawaiian-language newspapers were published, with something over 50,000 surviving pages. Together with Hawaiian-language books and archival materials, the surviving Hawaiian-language items, totalling perhaps a quarter-billion words, far outnumber all surviving materials from all the other indigenous languages within the United States.

Getting the newspapers (and, separately, the other materials) online was recognized as a priority by Ulukau, the Hawaiian Internet library (founded 2003). The newspaper collection began in August of that year as a draft website, announced verbally the following month to 11 people who were allowed to use word-of-mouth, but not to release the news to the media until further work had been completed on it. Usage that month, while the site was still in its beta version, showed 4,771 page views.

It was unclear at the time if this was a high or low number. There was not much of a baseline to compare the page views to as little had been done before with online access to Hawaiian-language content. The main intention at the beginning of the project was to provide access to the Hawaiian-language resources via a stable, well-designed website in conjunction with the affiliated sites that made up the library. To some degree it was done on the principle that the material should be made available and the hope was simply that it would work and be accessible to the people.

The advisory body to Ulukau had about 40 members, counting the 11 who attended the early meeting. The advisers included key representatives from all major stakeholders in the Hawaiian language community including archivists and educators. Word of mouth steadily raised usage numbers. In March 2004, when the newspaper collection went public (nupepa.org), the site spiked at 15,488 page views. Timed to coincide with the annual Native Hawaiian Education Convention, the site was officially released to the media. The new library was presented at the meeting, supported by an accompanying press release and an interview with a local reporter.

As luck would have it, the press liked the concept and a nice article ran on the front page of the Monday morning newspaper. In addition, a leading radio station picked up on the story and ran with it during their morning drive-time commute broadcast. Other radio stations and the evening TV news stations picked it up and reported the story. Word continued to spread.

Through July of 2004 the site exclusively used an English-language interface (menus and commands). In August the site launched a Hawaiian-language interface, and made that the default. This immediately proved popular and by the end of 2004, nearly two-thirds of usage was on the Hawaiian-language interface site. Usage that December totalled 18,370 page views with 63% of this traffic (11,625 page views) from the new Hawaiian-language interface.

Since then, usage continued to rise and for some years have held steady at 40,000-plus page views per month, with over two-thirds being on the Hawaiian-language interface.

The marketing lessons of this successful project are threefold.

First, have reliable digital-library software that is stable, solid, and simple. We use and highly recommend the absolutely fantastic Greenstone system, which continues to be developed by an academic team in Aotearoa (New Zealand) as a well-supported open-source system. We are also involved with a second-generation proprietary system, Veridian, that improves performance and allows for crowd sourced User Text Correction, and also recommend it highly. This system will replace our current collection sometime in the near future.

Second, customize your software in the direction of simplicity. From the beginning, this was the cornerstone of our integrated strategy. We built the Ulukau library from scratch with an eye towards the different elements working together in a simple, consistent manner. Our customizations of the digital library allowed for a very fast simple user experience with a minimum of clutter or other information.

Third, involve your stakeholders from the start. They are the ones who will be using it, and by being on the team from the earliest days, they buy in and spread the word. In our case, listening to our stakeholders created an environment where our marketing message kind of wrote itself.

The library had the intention of making available the archive of Hawaiian-language materials, the largest by far of any Native American language. The newspapers, in particular, had articles on news, but also genealogy, chants and stories about the culture. They were also valuable for word usage, uncovering words no longer used that were not preserved via dictionaries, and for evolving syntax. In addition to this rich content, the fact that so many newspapers still existed was newsworthy. The oldest, launched in 1834 is the oldest newspaper published west of the Mississippi.

We know that our patrons use the collection for a number of different reasons, but a few examples seem to represent the importance of providing access to Hawaiian-language books and archival materials.

Preserving and accessing history – Prior to the availability of the library there was no inventory of what pages had survived from Hawaiian-language newspapers. Finding coverage of historical events was a laborious process of going through numerous microfilm rolls for the various newspapers published at any one time. This task is now easily accomplished on the website by browsing through newspaper titles or dates, and by fast keyword searches.

Gaining access to Hawaiian-language teacher resources – Hawaiian-language teachers had been using Xerox copies of newspaper articles to help teach the language (and often the pages were copies of copies, many times over). The website opens up a wealth of resources and materials, all easily accessible online.

Preserving cultural heritage documents – I remember a teacher who attended a presentation of mine at a public-school gathering. She got up and shared a common misconception in her graduate school program that there was no written record of anything for Hawaiians. She recognized the library as a resource that now shows a literate people with dozens of historic Hawaiian-language newspapers. She was moved. This is also part of our on-going message.

## 3. MARKETING TEXT CORRECTION AT CALIFORNIA DIGITAL NEWSPAPER COLLECTION

The California Digital Newspaper Collection (CDNC) is the largest, freely-accessible archive of digitized California newspapers. The collection contains over 60,000 issues and nearly 550,000 pages—and growing, ranging from 1846 to the present. It is available for searching at http://cdnc.ucr.edu. The project is managed and hosted by the Center for Bibliographical Studies and Research (CBSR) at the University of California, Riverside. It has been supported in part both by the National Digital Newspaper Program (NDNP), a joint effort by the National Endowment for the Humanities and the Library of Congress, and by the Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. The CDNC has also partnered with local institutions around the state to digitize their newspapers and add the content to the archive.

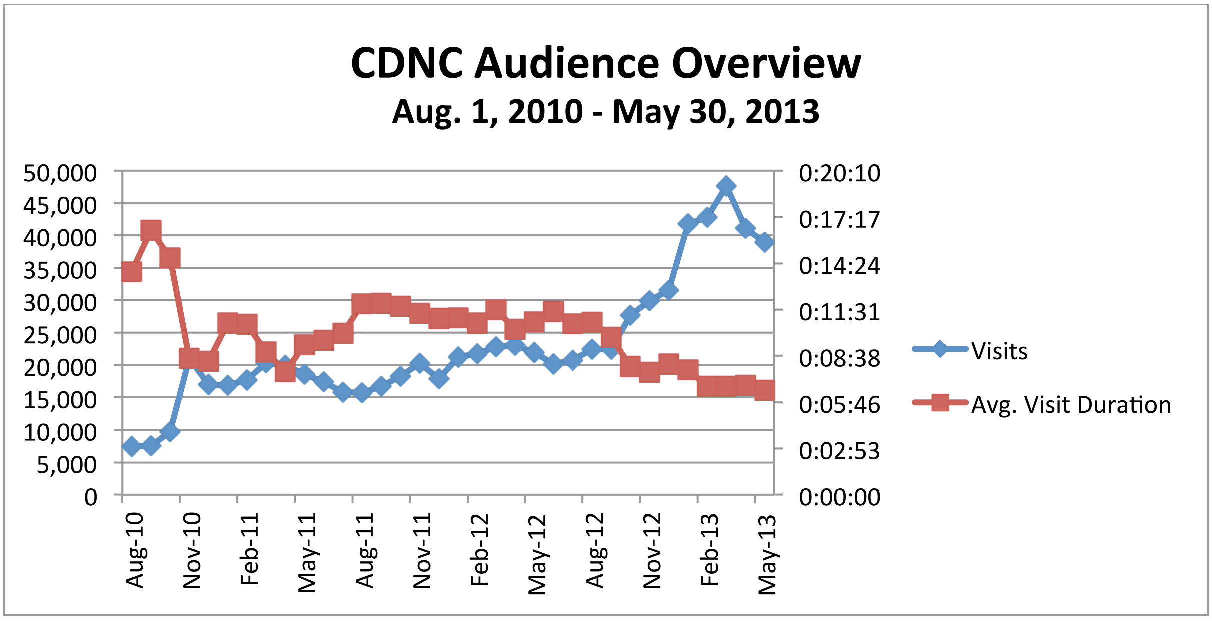
Work on digitizing California newspapers began in 2005, when the CBSR was selected as one of the first six participants in the NDNP. In October of 2007 the CDNC officially launched its website and hosted a conference in Riverside with participants and speakers from around the state attending. Much of the initial content for the archive came from work that was also submitted to the NDNP, but all pages were digitized to the article level rather than just the page, a practice the CDNC continues to this day. For the first two years of the project the CBSR developed and maintained its own hosting software, Zoopraxi; then in the fall of 2009 we decided to switch to Veridian, the hosting software ever since. In August of 2011 the CDNC, working closely with the developers of Veridian, enabled user text correction (UTC) within the archive, allowing users to register and then edit the computer-generated text. In the months since over 1300 individuals have registered, of whom nearly 600 have corrected over a million lines of text.

Staffing at the CDNC is small and time and resources are limited. Nonetheless, we have attempted to reach out to users and potential users through a variety of resources. Announcements about recent developments, for example the addition of new content, are always sent out on listservs and via email, posted on Facebook, sent to partner institutions like the State Library, local libraries, and the California Newspaper Publishers Association (CNPA), and when applicable, published as articles on the UCR homepage, articles that are frequently picked up and reprinted by the local media.

We know from surveys we've conducted that CDNC users tend to be older (80% were over 40 in the last survey) and are interested in genealogical research (60% in the same survey). We assume that "traditional" forms of communication, like email and listservs, are useful forms of outreach to this demographic. Once the results of a more recent survey are compiled, we will have a better sense of whether CDNC researchers also use social media like Facebook and Twitter. The CDNC has maintained a Facebook page for more than a year, and its use is probably best described as modest. The total number of "Likes" is now 286 and during any given week the number of people "Talking About" it ranges between 2 and 4. A search of Twitter reveals that on average over the last year there has been about one tweet per month on "California Digital Newspaper Collection," including tweets by our own staff.

The CDNC has maintained Google Analytics ever since the project started using Veridian and those numbers reveal some additional insights into outreach and use. The figure below depicts the number of visits versus average visit duration from August 2010 to May 2013. Monthly visits have increased noticeably twice since 2010, once at the very start and then again in the fall of 2012. The first increase coincided with the replacement of Zoopraxi with Veridian. The average number of monthly visits prior to September 2010 had been about 7,000. In November of that year it jumped to 21,000 and remained steadily around 20,000 for the next two years. At the same time, during the last quarter of 2010 the bounce rate increased from 19% to 60% and traffic from search engines increased from 30% to 60%. Indexing of the CDNC by Google clearly increased with the adoption of Veridian, which raised the number of visitors. The increase in visits at the end of 2012 is more difficult to explain. Between October 2012 and January 2013 the number of monthly visits increased to 41,000, yet the traffic from search engines remained steady at around 70% and the bounce rate at 60%. There was also no major announcement or outreach campaign by the CDNC during this period.

The statistics for average visit duration reveal the impact of UTC on CDNC use and suggest one reason for the increased number of visits at the end of 2012. In the fall of 2010 duration dropped precipitously from 16:28 to 8:30 minutes; time spent on the CDNC declined, not surprisingly, as indexing by Google and the bounce rate increased with the transition to Veridian. Between November 2010 and July 2011 the average visit duration fluctuated but never went above 10:41 minutes, averaging 9:14 minutes during those nine months. Then in August 2011, when UTC was introduced, the average duration jumped to 11:52 minutes. For the next year it remained fairly consistent, averaging 10:42 minutes. In September 2012 average duration fell to 9:48 and in October to 7:59 minutes. For the last nine months, through May 2013, it has averaged 7:23 minutes, the lowest in the project's history. Tellingly, by May 2013 the number of visitors coming from Google has increased to 26,000, up from 12,000 in July 2012. Combining these recent figures, one can conclude that absolute search engine traffic has increased recently, and though users are not "bouncing" off the CDNC, they are not finding what they expected and are leaving the site after only a few minutes.



There is surely much more one could uncover in the Google analytics. This brief survey suggests two efforts the CDNC can and will undertake soon to try to increase outreach and use. First, given the noticeable impact of UTC on the time users spend in the archive, we hope to develop other tools for researchers, such as tagging or commenting, that should encourage them to spend more time in the collection, while also enriching their experience and allowing them to contribute to the archive. Second, it would be interesting to see whether a concerted effort to announce developments with the project through social media would correspond to increased use of the CDNC, particularly traffic coming in through means other than search engines. CBSR staff will try to use social media more often and more consistently, and then assess through Google analytics what if any impact that effort has on use of the collection.

Lessons can also be learned through discussions with regular visitors and text correction users. Two of the top CDNC users share the reasons they visit and use the site, indicating some of the factors which motivate patrons to engage with DNCs and the community.

> *"I have always been interested in history, especially the development of the American West, and nothing brings it alive better than newspapers of the time. I believe them to be an invaluable source of knowledge for us and future generations."*
>
> *David, United Kingdom*

> *CDNC is an excellent source of information matching my personal interest in such topics as sea history, development of shipbuilding, clippers and other ships etc. ... Unfortunately, the quality of text ... is rather poor I'm afraid. This is why I started to do all corrections necessary for myself ... and to leave the corrected text for use of others. .... I am not doing this very regularly as this is just my hobby and pleasure.*
>
> *Jerzey, Poland*

## 4. THE SUCCESS OF A SOFT LAUNCH: THE CAMBRIDGE PUBLIC LIBRARY'S HISTORIC NEWSPAPER COLLECTION

Hidden deep within the pages of the September 28th, 1861 edition of the *Cambridge Chronicle* is a tiny note that the end of the world – postponed from 1843 – has been declared by the Millerites, a national, religious group predicting the Second Advent of Christ, will take place on Saturday October 12th. An article titled, "The Rights of Women!," published on August 10, 1848 is a tongue-in-cheek response to feminists Lucretia Mott and Elizabeth Cady Stanton, who had organized the Seneca Falls Convention a few weeks earlier. Page 11 of the May 24, 1890 issue of the *Cambridge Chronicle* lists a notice for a public meeting, hosted by the Indian Rights Association, promising the anticipated large audience perspectives from three "Indian students," a professor, and a clergyman. Each of these articles carries the weight of the political and social views from one of the oldest and largest cities in Massachusetts. This "old Cambridge news," previously hidden on rolls of microfilm, has tremendous value for researchers all over the world and is now a simple keyword search away from discovery.

The Cambridge Public Library, under the direction of the Archives and Special Collections began its digitization efforts by making accessible the historic Cambridge newspapers in its collection, including the *Cambridge Chronicle* – the oldest, continually published weekly in the United States. The library, located in Cambridge, Massachusetts on the other side of the

Charles River from Boston, exists in the center of the largest metropolitan hub north of New York City. Serving a population of 100,000 residents, the Cambridge Public Library system includes seven branches, circulates over 1 million books annually, and offers a variety of services that go well beyond the scope of a traditional library. The recently renovated main branch, where the Archives and Special Collections is located, receives over 1,000 visitors and circulates over 2,000 books daily. Often described as the "People's University," the library holds a prominent place in a community that is home to a rich history as one of the first settled towns in Colonial New England, and an intellectual history as Harvard University – the oldest institution of higher education in the United States - is just steps away from the main branch. Visitors flock to the Archives and Special Collections to delve into the local history: from historic newspapers to city directories and from photographs to personal papers. With funding from the Massachusetts Community Preservation Act, the Library digitized and made freely available all Cambridge newspapers in the public domain. Users are able to browse and search over 650,000 articles and view newspaper pages as they originally appeared for optimum historical context. For users to gain access to newspapers still in copyright, the Library digitized and made available over 54,000 newspaper subject and obituary cards that cover the years between 1950 and 2009.

The library chose Veridian as its software platform to manage the historic newspapers online not only for its ease of use and quick search turnaround, but also for its ability to overcome the technical challenges of delivering historical, digitized newspapers online by preserving the original newspaper layout. Veridian also allows users to become citizen archivists – a new approach employed by archives and libraries aimed at engaging the public around historical collections by asking anyone with an Internet connection to enhance or add to existing collections. Users fix textural errors created during the digitization process therefore immediately enhancing search results for all. People, places, and events once hidden by garbled text are now easily found. Those who make the most corrections have the honor of appearing in the text corrector "hall of fame," prominently displayed on the newspaper collection homepage. The Cambridge Public Library is the first public library in the United States using this software, and to date, over 40,000 lines of text have been corrected, creating a community of users and improving the database for all.

Since the Historic Cambridge Newspaper Collection (http://cambridge.dlconsulting.com ) went live in March 2012, the use rate has been steadily climbing. Over 22,000 users from around the world have visited the collection, viewing over 240,000 pages. To date, 40,000 lines of text have been corrected. Numbers continue to rise rapidly; as of May 2013, a daily average of 300 users visit the site, correcting 150 lines of text – which is amazing considering that just a year ago, patrons would have had to visit the Library and view the newspapers on microfilm. To put these figures in perspective, there have never been more than five patrons using the newspaper microfilm in one day. These statistics are remarkable for one major reason: the library has not yet launched a major marketing campaign around the collection.
Like many archives and special collections, the library has one full-time archivist overseeing all functions – from curating new materials to managing the special collections reading room of a large, urban public library. Given the time constraints of the archivist and counting on the knowledge of the National Library of Australia's Trove - the hugely successful user interactive newspaper database - the library decided to manage a soft launch of the Historic Newspaper Collection with a coordinated marketing effort to begin in the fall of 2013. The library's marketing strategy has been to introduce the collection to patrons, list the collection on the library's website, and allow for word-of-mouth marketing to spread to academics, genealogists, and local history buffs. The archivist has presented the collection at regional

professional conferences, and blogged about it and specific articles to draw more users worldwide. This word-of-mouth marketing strategy has allowed the library to refine its future marketing initiative, *Digitize This!*, a key element of which will revolve around empowering the citizen archivist to help preserve Cambridge's unique history. *Digitize This!* will include a redesign of the Archives and Special Collections website and blog, an e-marketing campaign, and a targeted marketing campaign geared towards walk-in library patrons using posters, brochures, bookmarks, and a text correcting contest. The library also plans to collaborate with the *Cambridge Chronicle* to highlight a column from the newspaper each week.

The success of the Historic Cambridge Newspaper Collection's soft launch illustrates how much of a need there is for "old news" to be made freely available to researchers. There are countless communities of users eager to explore, reveal, and celebrate local history – offer them a free, online, user friendly historic newspaper collection, and they will find and use it.

To collect data about the interactive nature of crowd sourcing, the library reached out to the patrons who are among the most frequent users of the online collection. Nearly all the responses replied that text correcting was easy, fun, and a public service – a way to enhance the database for everyone to enjoy and use. "I like to correct text because I find it fun to see how people thought and wrote about different subjects 'in those days,'" writes one patron in response to the survey. "I like to think of a subject, say tornadoes for example, and then, randomly choose the articles I will correct. If I find an article that I think is interesting, I almost feel like it is my obligation to correct it so that the next person that stumbles upon it will be able to read it without any problems. I also find it kind of relaxing to correct the text and very satisfying once I have finished a whole article." Similarly, the library's top text corrector responded with the following:

> *As an amateur historical researcher my time for research is very limited. Making time to travel to archives, libraries, and historical societies does not happen as often as I would like. The Cambridge Public Library's online newspaper collection has been an invaluable resource and it is fun. I am very grateful for all the help I have received over the years from so many research organizations. Correcting text has several benefits. It makes it much more likely that I will find a story if I decide to search for it in the future. It is a way of saying 'thank you' to the Cambridge Library for having such a great resource available and maybe I can make the next person's research a little easier. It is my own little historical preservation project.*
>
> *Daniel, Somerville, Massachusetts, USA*

When asked about any serendipitous historical discoveries through text correction, the patron explained that he discovered an ancestor he did know he had: "My Great-great grand parents had a daughter that died around her first birth day. She had become lost on our family tree." Because the print was small and the OCR garbled, the patron's relative never appeared in the search results. "If somebody else had fixed it earlier I would have been able to save several phone calls," explains the patron. "So it is a good example of why people who do already have information on a project, should not quit just because a story they expect to be there does not show up on a word search." The immediate benefit to all researchers is apparent to those who have corrected text. Crowd sourcing creates a community of users – natural citizen archivists eager to help each other and to share their knowledge. This community was an unexpected benefit to the implementation of the text correction module.

## 5. MARKETING VASSAR STUDENT NEWSPAPERS

Vassar College, in Poughkeepsie, NY, is a highly selective, coeducational, independent, residential, liberal arts college founded in 1861 and located approximately 75 miles north of New York City. With approximately 2,400 full-time students and more than 290 faculty members, Vassar is consistently ranked among the top liberal arts colleges in the United States. It is renowned for its pioneering achievements in education and its long history of curricular innovation (http://www.vassar.edu/about/). The Libraries collect materials and create digital collections with this tradition at the forefront, focusing on the breadth and diversity of the curriculum in close cooperation with the faculty as well as the chronicling of Vassar's history; the approach has resulted in more than one million volumes, including more than 50,000 rare book volumes and 500 archival and manuscript collections, in the library system.

In keeping with this tradition, the collection of 19th and 20th century student newspapers in the Vassar College Archives is a rich resource for information about the history of the school. Articles touch on a wide range of topics, and document ideas, issues, and events happening on- and off-campus. For many years, however, access was limited to browsing either hard copies or microfilm on site. As across-the-country digitization of collections became more widespread, we monitored how other institutions created digital versions of their school newspapers. We believed strongly that a digital version of the Vassar newspaper holdings should feature robust searching and other facets that would increase its functionality, and this meant that funding was needed. We shared our plans with College administrators, and eventually an anonymous donor came forward to support the project. We were ready to move forward with expanding access to the materials.

Thus, in 2011, Vassar College Libraries began the process to systematically digitize this student newspaper collection. Encompassing the back files of the newspapers from April 1872 to the present, our digitization efforts resulted in 4,287 issues available online, containing more than 55,000 pages and 165,000 articles. Using Veridian software through DL Consulting (http://veridiansoftware.com/), the Vassar College Libraries Digital Newspaper Archives (http://newspaperarchives.vassar.edu) launched in September 2011 with such notable features as a sophisticated search engine (including fuzzy matching, article classification, and date range), a "browse by date" feature for comparative analysis, and the ability to instantly share articles via social media. Given this rich feature set in the virtual collection and continuous requests for these items in the physical collection, we were confident that our online archives would be in high demand.

Instead, we discovered that our online collection had virtually no usage once launched. This was puzzling, given the nature of the site and the consistent interest in the content. As we began to analyze the problem, we found some peculiarities in the way our site was being indexed. Using Google Analytics to measure our site traffic, it was clear that users could not find our site easily: traffic generally came from the Vassar Libraries' website and never from "organic" searches or non-Vassar referrals. Three possible scenarios were explored. First, we thought perhaps that search engines were confused by our setup: because we were also sharing our newspaper data with a consortium group, the Hudson River Valley Heritage Historical Newspapers Collection (http://news.hrvh.org/), perhaps a search engine like Google interpreted the Vassar site as a mirror for this larger site. Second, we knew that any new site might take some time before it is properly crawled: maybe our site needed a few months to gain in popularity. Finally, we needed to examine our server setup. Was a robots.txt file, HTTP status report, or similar directives hindering a crawl of our site?

We discovered fairly quickly that our server's robots.txt file had restricted traffic too severely; search engines that respected this file steered clear of our site. While this information was powerful (and eliminated the concern that perhaps our site was misinterpreted as a mirror), and helped us correct the problem of a basic search engine crawl, it did not address a key component of our underlying problem: we wanted users to find us *as easily as possible*, making search engine optimization (SEO) the true solution for us. With over 54,000 pages of content, the Newspaper Archives posed a significant challenge for a search engine crawl – there was simply too much data to examine without the proper technical parameters in place. In other words, a crawler like Googlebot needed better directives to digest and index our content; for Google, this information is best expressed as an XML file called a Sitemap (http://support.google.com/webmasters/bin/answer.py?hl=en&answer=156184). Working with the team at DL Consulting, we optimized our site by setting up this specialized file, giving crawlers the information they need in a streamlined way, making any site indexing accurate and effective.

While we attacked the technical problem, we also began a marketing strategy to ensure that users could find our site while we worked to make search engines find us as well. In summer 2012, the Libraries formed an Outreach Committee to focus on new publishing outlets, and we created a blog, a Facebook page, and a Twitter account as a result. This provided a new opportunity to link to our newspaper content and reach this newly cultivated audience of current and former students (as well as other interested people). Using Facebook and Twitter, our college archivist began a "*Yesterday's News Today*" series, bringing greater patron engagement to our site. Approximately once per month, she posted links to newspaper articles featuring news events that occurred on a particular day, described a trend in Vassar's history, etc. This resulted in a noticeable uptick in traffic, essentially increasing our referral traffic from Facebook and Twitter. Meanwhile, our Digital Initiatives department worked with our Technical Services department to ensure that every newspaper title had a MARC record that linked directly to the title's digital presence. This MARC record was used in our library catalog and discovery service (Serial Solutions' Summon tool). Similarly, referral traffic spiked from these sites.

The impact of these strategies was dramatic: in the three months prior to indexing enhancements, total site visits were 447 (or about 5 per day). Three months post-enhancements, total site visits jumped to 12,522 – about 140 per day. A Sitemap and new traffic from our catalog, discovery service, and social media outlets provided an incredibly effective marketing strategy: search engines could find our site, links between services strengthened, and traffic began to flow to the site. Yet the effect of optimization cannot be understated. It is clear from our data that, while our marketing efforts resulted in new visits and visitors, our users most often found us through Google. Ensuring that our content was "Google-ready" meant an immediate impact on our traffic; creating easily digestible content for this search engine provided frequent first- or second-page results, increasing the likelihood that a user could find our content.

Our strategy continues to yield great rewards for our site. For example, from January 1 – April 30, 2013, our site saw more than 33,000 visits, or more than 275 per day. Traffic continues to increase, introducing more new visitors each month from organic searches; a nice side effect is that, as our site became discoverable through a search engine like Google, other sites began to link to our articles – so we now have referring traffic from non-Vassar websites. This new strategy, coupled with our continuing social media efforts, has made our Newspaper Archives one of the most-used sites that Vassar College provides.

## 6. CONCLUSION

Marketing your historical digital newspaper collection does not have to be an all-consuming, resource depleting exercise. There are a number of things you can easily do to ensure that your collection is discoverable and engaging for the patrons who visit.

The California Digital Newspaper Collection and the Cambridge Public Library demonstrated the importance of getting to know who is in your audience. An understanding of the demographics of your patron community will tell you a lot about the reasons they visit your collection, what they look for when they arrive, and how to keep them actively engaged so that they come back again and again, and invite colleagues and friends to visit as well. CDNC and the Cambridge Public Library facilitate this type of engagement by featuring User Text Correction in their collections. These collections have an established and growing number of registered patrons who regularly participate in correcting article text which not only helps to improve the quality of the collections, but contributes an aspect of community building and knowledge sharing. And as Trevor Owen's says in his blog[10]

> *"in addition to increasing search accuracy or lowering the costs of document transcription, crowdsourcing is the single greatest advancement in getting people using and interacting with library collections"*

The marketing lessons learned during the creation and launch of the Hoʻolaupaʻi Hawaiian Nūpepa Collection involved gathering a small group of stakeholders from the beginning of their project. Working with this group of people ensured buy-in from key representatives of their patron community, sparked a very successful word-of-mouth marketing initiative, and helped the library craft a marketing message that they later used for official announcements to the media.

The Vassar College Libraries Digital Newspaper Archives struggled initially with unexpectedly low usage rates after the launch in September 2011. Together with DL Consulting they worked to optimize the site for Google Indexing by setting up a specialized site-map to make it easier for search engine crawlers to find the collection. Three months after the site enhancement, total site visits jumped from 447 to 12,522 and traffic continues to increase with more new users arriving each month from organic searches. This is a simple, yet vital step which ensures that your collection is not invisible to search engines. The effort can result in a very large increase in daily visits, providing a solid foundation from which to build up the community through additional marketing and communication efforts.

[10] Paraphrased from Trevor Owen's blog http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/ (accessed June 2013).

The success of a digital collection goes beyond building and displaying it online. It is our hope that the information presented in this paper provides useful strategies and tactics that you can immediately apply to the marketing and promotion of your digital collection.

**User Survey for California Digital Newspapers Collection Survey**

The California Digital Newspaper Collection (CDNC) has been online since 2007. The purpose of this survey is to learn about the users of the website, for what purposes the collection is used, and to discover how many users correct text. We hope that you will complete the survey and become a regular at CDNC. Watch for additional new features to the site in the near future.

1. Do you use CDNC for work-related research or for personal purposes or for both?
□ Work
□ Personal
□ Both
□ Don't use the collection

2. Do you consider yourself a genealogist or family historian?
□ Yes
□ No

3. What are the main reasons you use CDNC (check all that apply)?
□ Genealogy or family history research
□ Community history
□ California history
□ Regional history of the West
□ General historical research
□ Other (please specify)

4. What type of information do you search for (check all the apply)?
□ Birth announcements
□ Wedding announcements
□ Death announcements and obituaries
□ Biographical information
□ General community history
□ Legal or court notices
□ Advertisements
□ Other (please specify)

5. Do you participate in any online genealogy forums?
□ FamilySearch
□ Ancestry.com
□ GenForum
□ MyHeritage
□ RootsChat
□ RootsWeb
□ Do not participate in any forums

□ Other (please specify)

6. Approximately how often do you visit CDNC?
□ Daily
□ Weekly
□ Monthly
□ Never

7. For a typical visit, estimate the number of minutes you spend on the CDNC website?
____ Minutes

Text at the CDNC website is computer generated using optical character recognition (OCR). The tet has many errors and consequently search results are less than perfect.

The website has a text correction feature that anyone can use. Corrected text is searchable by other users and gradually, through user contributions, the accuracy for website users will improve.

8. Do you currently use CDNC's text correction?
□ Yes
□ No

9. If you didn't know about CDNC text correction before taking this survey, will you try text correction in future?
□ Yes
□ No
□ Maybe

10. Do you have a social networking account?
□ No social network account
□ Delicious
□ Facebook
□ Google+
□ Twitter
□ Other (please specify)

11. Have you ever shared an artice or information you found in CDNC via social media such as Twitter or Facebook?
□ Yes
□ No

If you answered yes to the previous question, please briefly explain what you shared and which social media account you used. ________________________________________

12. We hope to add new features to the Veridian software used to host the CDNC. Which of the following features would you be likely to use? (Answer with **0** for wouldn't use, answer with **1** for might use, or answer with **2** for would definitely use.)

____ Download results from a search in spreadsheet of database format

____ Apply "tags" to an article for future recovery or clarification of its content
____ Write comments about an article
____ Download a high quality image of a page
____ Interact with other users on a CDNC forum
____ Save search results, articles, or pages to a personal collection
____ Add outside references to an article, for example, from Wikipedia or another website
____ Add a "georeference" to an article by linking it to a place on a map allowing users to see collections of articles about a specific place
____ Describe any other features you would like to be added to CDNC

13. Please give some basic demographic information about yourself.
____ State / Province
____ ZIP / Postal Code
____ Country

14. What is your age?
□ Under 20
□ 20 to 29
□ 30 to 39
□ 40 to 49
□ 50 to 59
□ 60 to 69
□ 70+

If you would not mind being contacted for further questions about your use of CDNC and about improvements or additional features that you would like, please give your name and email address and/or telephone number. This information will be held in strictest confidentiality and will by used by CDNC only for the aforementioned purposes. And of course you are always welcome to contact CDNC with questions and requests at cdnc@cbsr.ucr.edu.

__________ Name
__________ Country
__________ Email address
__________ Phone number