The search queries that took Australian Internet users to Wikipedia

Vivienne Waller
Institute for Social Research, Swinburne University of Technology, Mail HS3, PO Box 218, Hawthorn, Victoria 3122, Australia

Abstract

Introduction. This exploratory study analyses the content of the search queries that led Australian Internet users from a search engine to a Wikipedia entry.
Method. The study used transaction logs from Hitwise that matched search queries with data on the lifestyle of the searcher. A total sample of 1760 search terms, stratified by search term frequency and lifestyle, was drawn.
Analysis. Each search term was coded to indicate the subject of the query and weighted according to its position in the long tail distribution. Quantitative analysis was carried out using the statistical package SPSS.
Results. The results of the study suggest that Wikipedia is used more for lighter topics than for those of a more academic or serious nature. Significant differences among the various lifestyle segments were observed in the use of Wikipedia for queries on popular culture, cultural practice and science.
Conclusions. The analysis provides some analytical purchase on the complex nature of information search and the difficulties inherent in assuming a valid distinction between information search and entertainment. It is suggested that the term leisure search be used to identify information search that is in itself a leisure activity and not a search for particular information.

Introduction

Launched in 2001, Wikipedia describes itself as The Free Encyclopedia. Articles are written by an ongoing collaborative process and anyone can contribute to almost any article. As of April 2011, Wikipedia contained more than 3.6 million articles in English and more than 100,000 articles in each of 36 languages. (Source: Wikipedia home page at www.Wikipedia.org, accessed 22 April.)
According to Spinellis and Louridas (2008) *Wikipedia's* rate of growth is sustainable, meaning that *Wikipedia* is viable in the longer term as an online encyclopaedia.

**Comparing *Wikipedia* with traditional encyclopaedias**

Much research has compared *Wikipedia* with traditional encyclopaedias in terms of the quality of articles. There has also been a small amount of research on the coverage of topics where it has been found that *Wikipedia* has generally good coverage of the sciences, excluding medicine. The coverage of social science topics, especially law, appears to be less extensive (Halavais and Lackaff 2008). As may be expected, given the instant nature of publishing on *Wikipedia*, it also has greater coverage of more recent topics than do traditional encylopedias (Royal and Kapila 2009), especially popular music and recent fiction (Halavais and Lackaff 2008).

One of the basic tenets of *Wikipedia* is that all articles should present a neutral point of view. In theory, this, and the fact that anybody can edit a *Wikipedia* entry, should allow for alternative points of view to find expression. Hansen et al. idealise the *Wikipedia* model as ‘approximating the conditions for a Habermasian rational discourse’ (2009: 53), unshackling knowledge from traditional power relations. It appears, however, that the reality of *Wikipedia*’s entries is not so emancipatory. An empirical study by Elvebakk (2008) comparing *Wikipedia* entries on twentieth century philosophers with entries from online encyclopaedias written by academics concluded that *Wikipedia* does not generally present an alternative point of view to dominant Western thinking and values.

There is no consensus on the quality of articles in *Wikipedia*. Some commentators have made deductions about the quality of articles on the basis of the nature of the *Wikipedia* model of collaborative authorship. For example, the fact that anybody can contribute to an article and no one has overall editorial responsibility has led some to doubt the accuracy of articles (Gorman 2007; Keen 2007) or their coherence (Duguid 2006). Others draw the opposite conclusion, arguing that it is the hierarchy of user levels combined with the system of user-created policies and guidelines that ensures the quality of *Wikipedia* articles (Lipczynska 2005; McGrady 2009).

Empirical analyses of the content of actual *Wikipedia* articles have also yielded conflicting results. A study in *Nature* (Giles 2005; Giles 2006) found that a sample of *Wikipedia*’s science articles were as accurate as related entries in the *Encyclopaedia Britannica*. Chesney (2006) arranged for 258 academics to each read two *Wikipedia* articles, one within and one outside the academic’s area of expertise. Chesney found that the accuracy of information in the *Wikipedia* articles was high but that a given article was more likely to be rated as credible by experts than by non-experts. Ehmann et al. (2008) found that the quality of articles varied across disciplines.

Some commentators have suggested that the quality of *Wikipedia* articles could be improved by involving academics and librarians as active contributors to *Wikipedia* (Badke 2008; Pressley and McCallum 2008) while Lally (2007) suggests that *Wikipedia* should be used to extend digital collections.

**Need for research on how *Wikipedia* is actually used**

The relevance of research on the quality and coverage of *Wikipedia* and of debates concerning the role of academics and librarians hinges on how *Wikipedia* is actually being used. This is an area in which there is a paucity of research even though the popularity of *Wikipedia* is well established. In America, just over one in three adults use *Wikipedia*, with almost one in ten (8%) Internet users using *Wikipedia* on any day (Rainie and Tancer 2007). In Australia, in April 2009, *Wikipedia* was the twelfth most visited site on the Internet according to unpublished data from Hitwise Australia. In America, the likelihood of use increases with income and decreases with age, with those who have a university degree most likely to use *Wikipedia* (Rainie and Tancer 2007). In a US study by
Lim (2009), one third of university students reported that they use Wikipedia for academic purposes, generally for finding background information.

While it is established that Wikipedia is well used, there is less research on what it is used for, with most existing research based on surveys or interviews (for example, Luyt et al. 2008; Lim 2009). However, for a variety of reasons (such as poor recall, or wanting to impress the interviewer) how people say they use Wikipedia may not accurately reflect what they actually do.

In addition, most empirical analyses of Wikipedia articles or their use focus on academic content or context. For example, Lim (2009), Eisenberg (2009) and Luyt et al. (2008) all surveyed students. The debate in Nature (Giles 2005; Giles 2006) and the popular press about whether Encyclopaedia Britannica or Wikipedia is more accurate (e.g. Wikipedia study fatally flawed 2006) has focussed on scientific articles. The studies by Ehmann et al. (2008), Chesney (2006) and Elvebakk (2008) have all focussed on articles on traditional academic subjects. An analysis of the 100 most visited Wikipedia pages, however, found that more than half of these pages were related to sexuality and entertainment, particularly media celebrities and TV shows (Spoerri 2007). This finding points to the need to analyse more closely how a range of people, rather than just students, are actually using Wikipedia. Hence, this study addresses the following research questions:

• What type of information is actually being accessed on Wikipedia?
• Does use of Wikipedia vary across different segments of the online population?

This paper is based on an analysis of the content of the search queries that led Internet users from a search engine to a Wikipedia entry. It is located at the intersection of research into Wikipedia use and research into Web search. It follows in the tradition of the transaction log analysis (Jansen 2006), an established method in Web search studies. However, it differs from existing transaction log analyses (for example, Spink et al. 2001, Jansen and Spink 2006, Park 2009) in two important respects. Firstly, the sample has been selected and weighted to take account of the distribution of search queries. Secondly, data on the lifestyle of the searcher is attached to each query. While there have been previous studies that relate Web search to individual characteristics (for example, Ford et al. 2008), such studies been limited in sample size as they rely on interviews. In addition, the searches have been conducted in a controlled, and hence artificial, environment. The analysis reported in this paper is therefore distinctive because data on the lifestyle of the searcher have been attached to transaction logs.

Web search studies have been informed by models of information-seeking behaviour that assume that an information search is conducted in response to an information need (Spink and Jansen 2004, Knight and Spink 2008). However, Broder (2002) and Rose and Levinson (2004) have each demonstrated that a search engine user may not necessarily be searching for information. Broder (2002) introduced a taxonomy of Web search that added navigational searches and transactional searches to informational search. Navigational searches are those searches undertaken in order to reach a particular website that the user has in mind, while transactional searches are those searches undertaken in order to be able to complete a transaction, such as banking, shopping, or downloading files. Rose and Levinson (2004) preferred to replace Broder’s category of transactional search with the broader category of resource search, where the user's goal is to access an online resource other than information.

Before undertaking this study, it was assumed that the goal of users accessing Wikipedia was to locate information on a particular topic. Because Wikipedia is described as an encyclopaedia, this seemed a reasonable assumption. However, as discussed in the conclusion, the findings indicate that searches that lead to Wikipedia may not be searches for information.

It has been noted that, in order to improve search engine design, it is necessary to understand a user's intention or goal when conducting a Web search (Jansen et al. 2008, Rose and Levinson 2004). Similarly, it could be argued that understanding more about who is accessing what type of
The search queries that took Australian Internet users to Wikipedia Page 4 of 19

information on *Wikipedia* has implications for the design of *Wikipedia* pages and their accessibility through search engines.

For a user's search to appear in the sample, all of the following four conditions shown in Figure 1 needed to have been satisfied.

![Figure 1: Arriving at Wikipedia through a search engine](http://informationr.net/ir/16-2/paper476.html 23/06/2011)

The distribution of subjects looked up in *Wikipedia* is not an indication of subjects looked up on the Internet because the extent to which each of these four conditions is satisfied may vary across subjects. For some subjects, the user may be more likely to go directly to a Web page, or the relevant *Wikipedia* entries may be less likely to appear high in the results or the user may be less likely to choose *Wikipedia* from the search results. So, for example, while studies of Internet search queries found that more than 10% of queries related to computers or the Internet (Jansen and Spink 2006), only 5% of queries that led to *Wikipedia* were about computers or the Internet.

The following discussion summarises existing research in the United States, United Kingdom and Australia relevant to the likelihood that each of these conditions is satisfied.

**Condition 1: Use of Internet to find information**

It is evident that the Internet has become a major source of information, particularly because of its convenience (De Rosa *et al.* 2005; Horrigan 2006). A study taken by the Pew Internet and American Life Project in 2007 (Wells 2008) found that almost 60% of respondents would consult the Internet when they had problems they needed to address. In Britain, two-thirds of people turn to the Internet first when looking for information relating to a professional, school or personal project (Dutton *et al.* 2009). In Australia, where approximately 80% of the population aged over 15 use the Internet (Ewing *et al.* 2008), just over two-thirds (68.5%) of surveyed Internet users described the Internet as an ‘important’ or ‘very important’ source of information (Ewing *et al.* 2008).

There is not a great deal of data on what types of information the Internet is used for. The British Oxford Internet Survey asks people whether they use the Internet for getting information about local events, news, health, sports, jobs, humorous content or making travel plans. Two-thirds of Internet users in Britain (Dutton and Helsper 2007; Dutton *et al.* 2009), and approximately 80% of Internet users in the US (Fox and Jones 2009) look for health information online. In 2006, in the US, more than half (54%) had used the Internet to find out news or information about science (Horrigan 2006).

**Condition 2: Use of a search engine**

Little is known about the likelihood that a user uses a search engine rather than directly to a Web page for a particular subject. However, there is evidence of high overall use of search engines. In the US, PEW found that in 2008 almost half of all Internet users used search engines on a typical day (Fallows 2008). In Britain, almost two thirds (64%) of Internet users mainly use search engines when they are looking for information online (Dutton *et al.* 2009). In Australia, one in eight Website visits in the four weeks ending April 2009 was to a search engine (Source: Unpublished data from Hitwise Australia).
**Condition 3: Wikipedia entry appears in the search results**

The requirement that the relevant Wikipedia entry should appear in the search results can be restated as the requirement that the relevant Wikipedia entry should appear high in the search results. This is because more than two-thirds (68%) of users do not go past the first page of results and 92% do not go past the first three pages of results (iProspect 2008). It has been shown, however, that Wikipedia pages are often very highly ranked in search engine results, partly because of the dense structure of links in Wikipedia articles (Rainie and Tancer 2007; Spoerri 2007; Tann and Sanderson 2009).

**Condition 4: User clicks on link to Wikipedia**

Little is known about when a user will click on a link to Wikipedia and when they choose a different type of site. However, research by Tann and Sanderson (2009) suggests that a search engine is sometimes used as a shortcut to the relevant Wikipedia entry. In their study of university students, they found that more students expected to end up using Wikipedia than went straight to Wikipedia. Other research provides evidence to suggest that the likelihood of trusting Wikipedia as a credible source depends on the user's knowledge of the topic. Lim (2009) found that users who are new to a topic are likely to underestimate the quality of the relevant Wikipedia article. This may lead to reluctance to use Wikipedia for more important topics.

The situation is quite different regarding variation in Wikipedia use across lifestyle groups. For a particular subject, this variation may reflect differences in the propensity to use the Internet for information, to use a search engine rather than go directly to a Web page or to choose Wikipedia from the search results. Condition 3, whether or not a link to Wikipedia appears in the search results is, of course, independent of lifestyle group.

**Methods**

Wikipedia use can be measured in terms of the number of visits to particular pages or in the context of people's information searches. The number of visits to particular pages is of interest but this measure may be artificially inflated by the repeat visits of zealous contributors.

In the context of information searches, the most common way that users arrive at Wikipedia is through clicking on search engine results. In April 2009, two thirds (66%) of visitors (in Australia) to Wikipedia came directly from a search engine (Source: Unpublished data from Hitwise Australia) and almost all (93%) of these came from Google. This paper analyses the search queries that took people to Wikipedia in the four weeks ending 25 April 2009, using data provided by web analytics company, Hitwise Australia.

Hitwise measures Internet use by collecting data directly from Internet service providers. It collects the log files of proxy cache servers covering more than one third of Australian Internet subscriptions, including homes, businesses, schools, universities, and libraries.

The advantage of using transaction logs for investigating the sort of the information people access on Wikipedia is that transaction logs enable the study of a large sample of users, are unobtrusive, and do not affect user behaviour. Because transaction logs record what people actually do, they overcome the limitations associated with relying on what people say they do. The main limitation for a study like this is that the topic in which the user was interested can only be imputed by the researcher on the basis of the search query. A more certain method would be to observe users conducting searches and interview them at this time. In addition, the data used in this study does not provide information on those cases where Wikipedia appeared near the top of the search results but was not visited or where Internet users went directly to a Web page for information rather than by a search engine.
This study is quite different from Spoerri’s (2007) study of the 100 most visited Wikipedia pages in that it includes the long tail in its analysis. The term long tail was brought into popular usage by Chris Anderson (2006). It is used to refer to a Zipf distribution whereby a few items account for a sizable proportion of the total and an enormous number of items (the long tail) each contribute a tiny proportion to the rest. For example, the top one hundred search terms that brought people to Wikipedia in April 2009 accounted for just 4% of visits through search engines; hundreds of thousands of search terms accounted for the other 96% of visits with most only accounting for one or two visits each. Analysing the long tail is laborious as it requires manual coding of large numbers of search terms. It is important here to make a distinction between search terms and search queries. A search query is an instance of a user typing a query into a search engine. A search term is what is typed into the search engine. It is estimated that more than 600,000 search queries in the Hitwise sample took a user to Wikipedia and that at least 400,000 of these queries used search terms that appeared only once. The sample analysed in this study was selected from an extract of 50,000 different search terms that took a user to Wikipedia in April 2009. The distribution of the number of queries associated with each search term was mapped and divided into seven groups based on known and estimated search term frequency, as shown in Figure 2.

![Graph showing long tail distribution of search queries to Wikipedia](image)

**Figure 2: Long tail distribution of search queries to Wikipedia**

As well as including the long tail in its analysis, another unique feature of this study is that it matches search query data with data on the lifestyle of the searcher. It does this by taking advantage of the Hitwise proprietary method to match Internet use with Mosaic lifestyle profiles. Mosaic Australia has divided the Australian population into eleven lifestyle groups on the basis of a range of lifestyle information including income, education, employment, housing, consumer habits and preferences, entertainment, media use, and attitudes to government. This information is sourced from more than 200 variables derived from a range of data sources. The eleven lifestyle groups are listed in Table 1 alongside the relevant tagline which summaries the salient features. (See [http://www.mosaicaustralia.com.au/](http://www.mosaicaustralia.com.au/) for more information on the method and the lifestyle profiles).

<table>
<thead>
<tr>
<th>Lifestyle Group</th>
<th>Tagline</th>
</tr>
</thead>
</table>

http://informationr.net/ir/16-2/paper476.html
<table>
<thead>
<tr>
<th>Lifestyle Group</th>
<th>Tagline</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Privileged Prosperity</td>
<td>The most affluent families in the most desirable locations</td>
</tr>
<tr>
<td>B Academic Achievers</td>
<td>Wealthy areas of educated professional households</td>
</tr>
<tr>
<td>C Young Ambition</td>
<td>Educated and high-earning young singles and sharers in the inner suburbs</td>
</tr>
<tr>
<td>D Pushing the Boundaries</td>
<td>Young families living in recent developments on the fringes of major cities</td>
</tr>
<tr>
<td>E Family Challenge</td>
<td>Mixed family forms with stretched budgets in outer suburbs</td>
</tr>
<tr>
<td>F Metro Multiculture</td>
<td>Medium to high density areas with much cultural diversity</td>
</tr>
<tr>
<td>G Learners and Earners</td>
<td>Students and professionals living in high density, lower cost suburbs</td>
</tr>
<tr>
<td>H Provincial Optimism</td>
<td>Anglo-Australian blue-collar families in provincial settlements</td>
</tr>
<tr>
<td>I Farming Stock</td>
<td>Rural landowners and workers in agricultural heartlands</td>
</tr>
<tr>
<td>J Suburban Subsistence</td>
<td>Low income, low-spending households in major regional and outer metro</td>
</tr>
<tr>
<td>K Community Disconnect</td>
<td>Older blue-collar workers and retirees in country and coastal locations</td>
</tr>
</tbody>
</table>

**Table 1: Lifestyle groups and their tagline**

These lifestyle groups have been derived for market research rather than academic research. However, this does not detract from the usefulness of the categories for a study like this which can be understood as an analysis of the consumption of the information product *Wikipedia*. For each lifestyle group, a random sample of 160 search terms, stratified by search query frequency, was selected without replacement. In this way, a total sample of 1,760 search terms was drawn. The sample was divided equally across the lifestyle groups to ensure the same precision for each lifestyle group.

The analysis weighted queries by search term frequency. To avoid the dominance of those search terms that were used most frequently, search queries were also weighted according to their position in the long tail distribution. For example, queries in the sample that used search terms that appeared more than 500 times were weighted to contribute a total of 1.7% of visits in the sample (see Figure 2). Reporting on the total frequencies for each subject required an additional weighting taking into account the representation of each lifestyle group in the online Australian population.

Each of the search terms in the final sample was examined and coded to indicate the subject of the search term. Each search term was looked up in Google to enable inspection of the relevant *Wikipedia* entry in the search results. Those terms which related to an Australian subject were also separately identified.

The starting point for the coding in this study was that developed in an analysis of the subject of library catalogue queries (*Waller 2009*). The Library of Congress classification was used by Halavais and Lackaff (*2008*) in their analysis of the topical coverage of *Wikipedia*. However, as *Park (2009)* explains, traditional classification systems such as this are not well suited for classifying Web queries. The categories introduced by *Wikipedia* in 2004 were used by Holloway *et al. (2007)* in their analysis of the topical coverage of *Wikipedia*. However, these categories were...
not suitable for this research either as they are not mutually exclusive and not sufficiently fine-grained. For example the top level category Culture and the Arts includes topics as diverse as sport, tourism, music, philosophy, TV, and food. (Portal: contents/categories n.d.).

Using subject codes similar to those of Waller (2009) will allow the subject of the library catalogue queries to be compared with the subject of Wikipedia queries and the subject of search engine queries in a subsequent analysis. Drawing from the grounded theory technique of open coding (Strauss and Corbin 1998) the coding scheme in Waller (2009) was generated from close examination of the data and the creation of codes that most closely described the content of the search queries. For this study of Wikipedia use, care was taken not to force the search terms into categories that did not reflect the substance of the term and additional codes were created as required. For example, the code pornography was not needed for the analysis of library catalogue queries, but was an appropriate code for some of the Wikipedia queries. The resulting fifty-two codes closely described the content of the search queries and these were amalgamated into the following twelve broad subject groupings.

- Popular culture (popular music, TV show, actor, movie, video game, celebrity)
- Cultural practice (sport, religion, cultural practice not elsewhere classified)
- Computing/Web
- Health
- History
- Science
- Place/building
- Contemporary issues
- Book/author
- High culture
- Other
- Unknown

In most cases the classification was straightforward and it was possible to code 98% of queries with a primary code. Where the meaning of the term was unknown or ambiguous, the code unknown was assigned. The presentation of the findings includes descriptions of each category to enable application of the coding scheme to other data.

There are a number of caveats. The coding inevitably involves some subjective decisions about the appropriate category in which to place a particular Website or term. Although the contribution of each Website or search term to the whole is small in the long tail, there is a degree of imprecision to the classifications. In interpreting the data, the focus should be on the overall pattern, rather than on the precise size of each category. In recognition of the imprecision, the percentage size of the categories is reported to the nearest integer.

Although the Internet service providers that supply data to Hitwise are a representative cross-section of such providers, there may be some sample bias, the direction of which is impossible to detect.

The segmentation of the Australian population is according to household rather than individual user. Some household types only include particular age groups within households, for example, retirees, or people in their twenties and thirties. Within households with families, it is not possible to distinguish the children’s use of Wikipedia from that of the adults.

It should also be borne in mind that Hitwise measures visits not visitors and so the data set may include a number of queries from the one individual. This is, however, unlikely, given the large size of the set of queries from which the sample was drawn.
Findings

Who were the visitors?

Figure 3: Lifestyle profile of visitors to Wikipedia, compared with their representation in the Australian population. (Data source: Hitwise)

Figure 3 compares the lifestyle profile of visitors to Wikipedia with the lifestyle profile of Australian households and the online Australian population in the research data. It can be seen that lifestyle group G (Students and professionals living in high density lower cost suburbs) are the smallest group, accounting for 4.8% of the Australian population and 4.3% of the online Australian population. Lifestyle groups E (medium to high density areas with much cultural diversity) and K (Mixed family forms with stretched budgets in outer suburbs) are the largest. These each account for 14.3% and 14.0% respectively of the online population. Comparing the representation of any particular lifestyle group in the Australian online population with their representation in visits to Wikipedia, it can be seen that the distribution of visits to Wikipedia approximates the actual distribution of the online population: the largest difference is less than two percentage points. However, the data indicates that amongst visitors to Wikipedia there was a slight over-representation of people who were better-off and had higher educational attainment and a slight under-representation of people who were socially or economically disadvantaged and who lived in rural or suburban fringe areas.

What were they searching for?

![Distribution of topics that took search engine users to Wikipedia]

Search queries that were obviously intended to lead to the Wikipedia site (*Wikipedia*, *wiki*, *Wikipedia encyclopaedia*, *site:en.Wikipedia.org*, and *www.Wikipedia.org*) were regarded as navigational queries (**Broder 2002**). Less than 2% of all queries were navigational. All other queries about a particular topic were regarded as informational queries, even if they also contained the word *wiki*.

A distinction was made in the coding between High Culture and Popular Culture. Drawing from Gans (**1999**), the code High Culture was used for queries relating to fine art and classical music while Popular Culture was used for queries relating to popular music, TV shows, actors, movies, video games or celebrities. While less than 2% of queries that took users to *Wikipedia* related to high culture, more than one third (36%) related to popular culture. As Figure 4 shows, popular culture queries most commonly related to popular music. Almost all of the queries tagged as popular culture were for names: the names of TV shows, actors, musicians, songs, movies or video games.

Only 15% of queries related to Cultural Practice. As can be seen from Figure 3, 5% were queries about sport (for example, *david beckham*, *figure skating*), 1% were about religion (for example, *buddhism*, *luke the evangelist*), and 9% were about cultural practices not elsewhere classified. This last group comprised queries about aspects of everyday life, including product names, foods, holidays and ceremonies (for example, *toyota corolla*, *jager bomb*, *green tea ritual*, *near death experiences*, *anzac day*) as well as words or expressions for which it seemed that the searcher was
seeking a definition or some background. Examples of this latter group include the words oxymoron and dessert. At the time of writing, the relevant Wikipedia page is the first Google search result for either of these two queries.

Queries about Science included queries about the biological sciences (for example, European honey bee, spectacled eider, biggest crocodile in the world) the physical sciences (for example, chlorine dioxide, black holes, transitional metals) and mathematics (for example, fibonacci sequence, pi, radial basis functions). Only 7% of queries were tagged as science.

Queries tagged as Health included queries relating to mental health (for example, psychosis symptoms, depression), sexual health (for example, herpes, urethritis), and psychology (for example, five stages of grief, nlp, resilience), drugs (for example, heroin, temazepam overdose) as well as other health queries (for example, rotavirus vaccine, peppermint tea benefits. Only 7% of queries related to health.

There were 6% of queries tagged as History. These were queries about historical events or periods (for example, mccarthyism, reasons for australia's colonisation), historical figures (for example, ned kelly, st Catherine of siena) and artefacts (for example ancient Egyptian tomb). One third of these queries related to military history (for example, vietnam war, battle of vimy ridge).

Contemporary issues, Computing and the Web and Book/author each accounted for 5% of queries. Contemporary issues was used for queries about current affairs (for example, fannie mae and freddie mac collapse, description of israeli and palestinian conflict), the environment (for example, effects of co2 emission, world oil usage per year), organisations (for example, dyno nobel portal), the law (for example, testimony, define:liquidated damages), political figures (for example, nelson mandela), and political or sociological concepts such as diaspora and pragmatics.

Computing or the Web included queries on computing software and hardware (for example, what is a cvbs input), gadgets (for example, qkphone 911 and sony walkman mp3 flash player e-series), Websites such as youtube, google and last fm, and Web tools (for example, what is irc or irc bots).

Book/Author queries were the names of works of fiction, comics, or writers, (for example, dave eggers, taming of the shrew, bleach manga) as well as other book-related queries (for example, poetry prizes).

The final 2% of queries tagged as Other are those topics too small to be reported on separately. These were all either queries about pornography, explicit sex, genealogy or business.

**What was each lifestyle group looking for?**

There were some distinct differences in what different segments of the online population were looking for on Wikipedia as Table 2 shows. The differences are significant (p<0.001) for queries relating to Popular culture, Cultural practice and Science.
The search queries that took Australian Internet users to Wikipedia Page 12 of 19

| Computing/Web | 6.3 | 4.1 | 2.1 | 6.6 | 2.6 | 6.5 | 6.2 | 3.3 | 5.8 | 8.2 | 6.1 |
| Place/building | 6.0 | 5.6 | 5.3 | 1.1 | 6.2 | 2.7 | 4.9 | 5.5 | 9.1 | 5.0 | 6.4 |
| Contemporary issues | 3.2 | 6.2 | 5.5 | 4.7 | 3.8 | 2.9 | 3.9 | 8.2 | 3.8 | 4.5 | 10.9 |
| Book/author | 7.2 | 4.8 | 4.1 | 4.4 | 9.0 | 3.2 | 3.8 | 4.1 | 4.9 | 6.7 | 1.1 |
| Wikipedia | 3.0 | 1.6 | 1.7 | 2.2 | 1.7 | 4.7 | 2.4 | 2.3 | 1.7 | 1.6 | 1.4 |
| High culture | 1.5 | 1.2 | 3.0 | - | 2.1 | 0.6 | 0.7 | 0.6 | 4.2 | 0.6 | 4.2 |
| Other | 0.6 | 1.1 | - | 3.0 | 0.6 | 0.6 | 1.3 | 3.7 | 2.5 | 1.6 | 4.1 |
| Unknown | 0.7 | 0.6 | 3.9 | 3.1 | - | - | - | 2.0 | 4.4 | 0.6 | 3.1 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 2: Comparison of search topics for the various lifestyle segments

Those most likely to go to Wikipedia for popular culture information were those from group F, medium to high density areas with much cultural diversity and those from group G, students and professionals living in high density, lower cost suburbs. Those least likely to go to Wikipedia for pop culture queries were those from group K, older blue-collar workers and retirees in country and coastal locations. The presence or absence of children or young people does not entirely explain the differences across the various lifestyle segments. For example lifestyle segment E (mixed family forms with stretched budgets in the outer suburbs) had a relatively small proportion of visits related to popular culture.

In general, those with higher income and living in high-density or wealthy areas were more likely to arrive at Wikipedia through pop culture queries than predominantly lower income Anglo-Australians living in the suburban fringe or country towns.

Those from older and lower income households in outer metropolitan, regional and country locations were the most likely to arrive at Wikipedia via cultural practice queries. These did not tend to be queries about sport but were more general queries about aspects of cultural practice including definitions or background on words or expressions. Those living in high-density areas with high cultural diversity were the least likely to arrive at Wikipedia through cultural practice queries.

Those most likely to arrive at Wikipedia through science queries were those from group B, wealthy areas of educated professional households and those from group E, mixed family forms with stretched budgets in the outer suburbs.
Only 7.3% of queries were about Australian topics. There was no significant difference in this across the lifestyle groups but, as would be expected, the proportion of Australian queries varied across topic areas. As shown in Figure 5, almost one in five (19%) queries about places or buildings were about Australian places or buildings and 13% of queries about history were about Australian history. One in six (16%) queries about cultural practice related to Australia. These were mainly queries about sport and Anzac Day (an Australian public holiday that occurred during the sampling period).

**Discussion**

**Use of Wikipedia**

As in the United States, those in Australia who have higher educational attainment and a higher income are slightly more likely to use Wikipedia.

When the coverage of articles is compared with the subject of articles being accessed, it is clear that certain subject areas are being more intensively accessed than others. According to Halavais and Lackaff’s (2008) study of the coverage of Wikipedia articles, the topics with the greatest number of articles on Wikipedia were general history and science, (each with approximately 13% of all articles) followed by geography, social science and literatures. Together, these five subjects account for 57% of articles. The analysis presented here suggests a very different pattern of use, whereby those five subjects accounted for 28% of visits. While articles about music, including
popular music, account for 7% of Wikipedia articles, queries about popular music accounted for double that proportion (14%) of visits.

The research by Lim (2009) and by Ehmann et al. (2008) provides clues as to why this may be so. Lim's research suggests that the average user may be wary of the credibility of Wikipedia for complex topics. Moreover, the research by Ehmann et al. (2008) suggests that most of the articles relating to traditional academic subjects, such as sociology, medicine, biology and philosophy, are difficult to read. In contrast to traditional academic subjects, information about popular culture and cultural practices is not conceptually difficult. Some researchers go so far as to classify queries lodged about favourite music and TV as entertainment rather than information-seeking (Ewing et al. 2008; Luyt et al. 2008). Together queries about popular culture and cultural practice accounted for just over half (51%) of all queries.

Previous research has found that, in terms of the content of the most popular search terms, Australia has similarities with the United States, Canada, Korea and the Netherlands (Segev and Ahtiv 2010). Further research is needed to gauge whether this similarity in content of the most popular search terms extends to a similarity in the type of subjects being accessed in Wikipedia via search engine queries.

The importance of including the long tail

This study shows that looking at only the most common search terms or the most popular Wikipedia pages can be misleading. It demonstrates the importance of analysing the long tail of search terms or visited pages when analysing the type of information accessed on Wikipedia. As already noted, in April 2009 the 100 most common search terms that took people to Wikipedia accounted for only 4% of all visits to Wikipedia from search engine results. There is no a priori reason to assume that the distribution of topics in the most common 100 search terms is the same as the distribution of topics along the long tail. There were some similarities to Spoerri's (2007) analysis of the top 100 pages, but also important differences. The proportion of queries relating to entertainment in the top 100 and along the long tail appears to be quite similar (Spoerri's code of Entertainment matched the code of Popular Culture grouped together with Fiction and Comics). Spoerri found that 43% of the 100 most visited Wikipedia pages were related to entertainment. In this study, the equivalent proportion of search queries (queries relating to popular culture plus queries relating to fiction and comics) along the long tail was 39%. However, the findings on queries relating to sexuality were quite different. While pages providing information on sexuality accounted for 10% of the most visited pages (Spoerri 2007), taking the long tail into account, queries on sexuality accounted for less than 1% of search queries that took people to Wikipedia.

Who uses Wikipedia and for what?

Wikipedia's appeal appears to be fairly even across different segments of the Australian population. However, as in the United States, it seems that those in Australia with more income are more likely to use Wikipedia and older people with lower income are less likely to use it. It is not clear from the data why this is so.

Significant differences across the lifestyle segments were observed in the proportion of queries about popular culture, cultural practice and science. It is possible that the observed differences between lifestyle groups in subjects of queries are to do with differences between lifestyle groups in any one of the conditions illustrated in Figure 1. In other words, it is difficult to know whether these differences are a reflection of different interests, a different propensity to use the Internet or a search engine to research these topics, or different levels of appreciation for the Wikipedia content on these topics.

Those most likely to arrive at Wikipedia through pop culture queries were those from medium to high density areas with much cultural diversity and students and professionals living in high
density, lower cost suburbs. Older blue-collar workers and retirees in country and coastal locations were much less likely to have queries relating to pop culture and much more likely to have queries relating to cultural practice. Although the explanation could be that this group is less likely to use the Internet for pop culture information, or they go directly to other Internet pages for pop culture information, this is unlikely. It is more likely that this group have less interest than other groups in pop culture. This explanation is borne out by the fact that according to Mosaic, this group identifies strongly with all things Australian (http://www.mosaicaustralia.com.au/). Although the proportion of specifically Australian queries was not significantly different for this group, the vast majority of the pop culture queries to Wikipedia (and arguably, the vast majority of pop culture) relate to specifically American products.

A US study (Horrigan 2006) found that people who turn to the Internet most for science information are more likely to be young and well educated. This study found that those most likely to arrive at Wikipedia via science queries included the well-educated (those from group B wealthy areas of educated professional households) as well as those with lower educational attainment (those from group E mixed family forms with stretched budgets in outer suburbs). It is not at all clear from the data why this is so.

**Conclusions**

This study contributes to studies on Wikipedia use and on the field of Web search.

It informs debates concerning the quality and coverage of Wikipedia and the possible role of academics and librarians by providing evidence of what parts of Wikipedia are actually being accessed. The existing controversy about the accuracy of articles in Wikipedia has focused on articles about serious academic subjects. Concerns about the accuracy of the information contained in Wikipedia may be partially allayed by an understanding of the type of information that people are mainly accessing. The current study suggests that it is the entries on popular culture that are being accessed most frequently. Half of the visits to Wikipedia from search engines are to find out information on popular culture or cultural practices. Nevertheless, it should also be borne in mind that even though the percentage of Wikipedia queries on serious subjects may be small, the numbers of people visiting Wikipedia are high. Indications are that amongst those Internet service providers sampled by Hitwise Australia, more than 1,000 people are using Wikipedia every day to search for information on contemporary issues and even more are searching for health information.

It appeared that a small proportion (approximately 2%) of visits were to find the meaning of words. It will be interesting to see whether this proportion increases over time and whether Wikipedia becomes significant as a dictionary as well as an encyclopaedia.

Significant differences among the various lifestyle segments were observed in the use of Wikipedia for queries on popular culture, cultural practice and science. This analysis of transaction logs does not provide enough information to explain these differences. Finer-grained research on where people go for particular types of information should enable a clearer understanding on the propensity of particular types of people to use the Internet for information on particular topics and to turn to Wikipedia for this information. It is also important that future research analyse search queries that are actually typed in to Wikipedia to see if there is any difference in the spread of topics accessed. Given that search engines account for two thirds of visits, it is unlikely that the main findings of this study would change.

The analysis provides some analytical purchase on the complex nature of information search and the difficulties inherent in assuming that there is always a meaningful distinction between information search and entertainment or leisure. While the common presumption is that Wikipedia is a source of information, the high proportion of popular culture queries suggests that for many,
Wikipedia could be considered a source of leisure. This highlights the need for a term to distinguish between information search where finding the information is the goal and information search that is an end in itself. I suggest the term leisure search to indicate when the searching for information is itself a leisure activity. This is very different from infotainment, which is the imparting of information through the medium of entertainment. With infotainment, what is important is that the information is imparted to the viewer or listener. With leisure search, the searching is itself a form of leisure and the question Did you find what you were looking for? does not make sense or is not relevant. The phenomenon of leisure search has existed before the Internet; for example, looking up particular terms in an encyclopaedia as a way of passing the time. However, whereas leisure search would have been quite an uncommon activity in the past, the ease of looking up anything on the Internet has facilitated the expansion of leisure search to become a very common activity. It is possibly similar in nature to surfing the web as a leisure activity.

Hence, this study contributes to the field of Web search user studies, indicating that use of a Web search engine may not be always be to achieve one of the goals identified by previous research. Rather than having a particular informational, navigational or transactional goal in mind, a search engine user may be engaging in a leisure activity. Whether an information search is leisure search or a search for information cannot be precisely determined by the subject. For one person, idly looking up science information is a leisure pursuit while for another, looking up a popular musician may be a search for a particular piece of information. It is reasonable to assume, however, that for most people a search related to a traditional academic subject will be an information-seeking activity while a query about popular culture is more likely to be to be a leisure activity than an information-seeking one. More research is needed to shed light on this issue. In the meantime, the implication for Web search user studies is that it is important to be specific about topics when asking people about, or reporting on, information search. In terms of improvement of search engine design, the implication is that it is important for search engine designers to understand more about the phenomenon of leisure search. This phenomenon could also have implications for the design of Wikipedia pages relating to popular culture and their accessibility via search engines.

Acknowledgements

This research was funded by Australian Research Council grant LP077215. The author would like to thank Hitwise for their generous support, in particular, Sandra Hanchard, Ben Carroll and Bill Tancer as well as the State Library of Victoria and Dr Denny Meyer. Thanks also go to the anonymous reviewers for their helpful comments.

About the author

Dr. Vivienne Waller is a Research Fellow at the Institute for Social Research at Swinburne University of Technology, Australia. She received Bachelor's degrees in both Science and the Arts from the University of Melbourne, Australia and her PhD from the Australian National University, Australia. She can be contacted at: vwaller@swin.edu.au.

References


• Keen, A. (2007). *The cult of the amateur: how today's Internet is killing our culture*. London: Nicholas Brealey.


