

Accessibility and Distribution of Information on the Web

[Steve Lawrence](#) and [Lee Giles](#)

This page summarizes our study in the journal *Nature*:

"Accessibility of information on the web", *Nature*, Vol. 400, pp. 107-109, 1999.

Send email: lawrence at necmail.com to request a copy (no cost). We would be interested to know where you heard about the study.

[Frequently Asked Questions](#)

[Previous Study](#)

[Publications](#)

[Home](#)

```
@Article{ lawrence99accessibility,  
  author = "Steve Lawrence and C. Lee Giles",  
  title = "Accessibility of Information on the Web",  
  journal = "Nature",  
  volume = "400",  
  pages = "107--109",  
  year = "1999"}
```

Search engine coverage has decreased	Search engine coverage relative to the estimated size of the publicly indexable web has decreased substantially since December 97, with no engine indexing more than about 16% of the estimated size of the publicly indexable web. (Note that many queries can be satisfied with a relatively small database).
Unequal access	Search engines are typically more likely to index sites that have more links to them (more 'popular' sites). They are also typically more likely to index US sites than non-US sites (AltaVista is an exception), and more likely to index commercial sites than educational sites.
Out of date	Indexing of new or modified pages by just one of the major search engines can take months.
Information distribution	83% of sites contain commercial content and 6% contain scientific or educational content. Only 1.5% of sites contain pornographic content.

800 million pages	The publicly indexable web contains an estimated 800 million pages as of February 1999, encompassing about 15 terabytes of information or about 6 terabytes of text after removing HTML tags, comments, and extra whitespace.
Low metadata use	The simple HTML "keywords" and "description" metatags are only used on the homepages of 34% of sites. Only 0.3% of sites use the Dublin Core metadata standard.

The web is transforming society, and the search engines are an important part of the process. The web and search engines represent a significant improvement for communication, providing efficient access to an increasing amount of information. However there are limitations to the current search engines, improvements to which may help to maximize the benefits of the web.

85% of users use search engines to find information (GVU survey). Consumers use search engines to locate and buy goods or to research many decisions (such as choosing a vacation destination, medical treatment or election vote). However, the search engines are currently lacking in comprehensive and timeliness, and do not index sites equally. The current state of search engines can be compared to a phone book which is updated irregularly, is biased toward listing more popular information, and has most of the pages ripped out.

Search engine indexing and ranking may have economic, social, political, and scientific effects. For example, indexing and ranking of online stores can substantially effect economic viability; delayed indexing of scientific research can lead to the duplication of work; and delayed or biased indexing may affect social or political decisions.

For more details request: lawrence at necmail.com a copy of the Nature article.