Ghana Library Journal, Vol. 20, No. 1, 2008

COVERAGE OF AFRICA-SPECIFIC SUBJECT AREAS AND SOURCES BY SOME MAJOR INTERNET SEARCH ENGINES

M. A. Tiamiyu and O.A. Salako
Africa Regional Centre for Information Science,
University of Ibadan, Ibadan, Nigeria.

ABSTRACT

The study compared the relative coverage of Africa- and Nigeria-specific sources of information on selected subjects by major Internet search engines. Ask, Google and Yahoo! were tested using Boolean queries comprising each of 20 highly Africa-pertinent subject key words and each of four regional terms Africa, Europe, Africa NOT South, and Nigeria. The quantities of hits by the engines for the same or different queries were then compared. The study found that Yahoo! tended to retrieve more hits on most of the subject areas than either Google or Ask; that the search engines differed often markedly in their ability to retrieve information on the different subject areas; and that South (or southern) Africa-specific information did not overly dominate the Africa-specific information retrieved by the engines. The findings show that more studies are required to test both the quantity and quality of sources retrievable through a greater diversity of search engines, and using more complex search queries containing narrower subjects and the names of individual African countries, regions, peoples and other entities. The study is expected to stimulate research on the use of specific Internet technologies, such as search engines, for Africa-specific purposes, such as the retrieval of Africa-specific information.

KEYWORDS: INTERNET, INFORMATION RETRIEVAL, INFORMATION SOURCES, INFORMATION ACCESS

Introduction

The storage and retrieval of information has always required the development of access mechanisms, whether for use by the information specialists who organize and manage the information, or by information end users. Until the development of electronic means of data storage, tools for accessing information were generally in the form of ordered lists, such as book or card catalogs, printed indexes, etc. With the advent of electronic databases and the Internet, the development of access mechanisms became more complex, requiring knowledge of database

structures and computer programming, as well as an understanding of how such systems would be used.

Computer programs for searching Internet resources began with fairly simple programs developed for specific uses: *Archie* to identify potentially relevant files for download; *Veronica* to identify potentially relevant menu items at Gopher sites; *Wide Area Information Service (WAIS)* to provide access beyond file names or menu titles to the actual contents of electronic documents. The advent and increasing growth of the World Wide Web (www) demanded expanded information indexing and search capabilities. The first of these more sophisticated search programs, *Lycos*, automatically identified words in document titles or file names to build indexes which were subsequently searched to retrieve the documents. Many other search programs (now known as search engines) have emerged since, some following the lead of Lycos by indexing document titles, others building on the WAIS tradition of limited full-text indexing, and still others, such as Yahoo! based on both human subject cataloging and automatic indexing techniques. The Internet now bubbles with a growing number of search engines, as well as *meta-search engines* (i.e. search engines that search multiple search engines and specialist subject directories for information) and *intelligent agents* (that scour the Internet for Web pages and information meeting specified criteria). Popular search engines, meta-search engines and directories today include AltaVista, Ask, Excite, Hotbot, Infoseek, Lycos, WebCrawler, WWW Worm, WWW Virtual Library and Yahoo!.

The term "search engine" encompasses a wide variety of Internet-based services which provide searchers with online access to Internet resources. The term is also often used generically to describe both crawler-based search engines and human-powered directories. These two types of search engines obtain their listings in different ways. Crawler-based search engines, such as Google, use computer programs to periodically "crawl" or "spider" through the Web to find web pages, and automatically extract/create and store descriptions of the pages in their databases. By contrast, human-powered directories such as the Open Directory, depends on human reviewers and indexers to write or evaluate submitted descriptions of web sites which are stored in their databases. Subsequently, in order to retrieve web pages in response to user queries, both crawler-based engines and human-powered directories use computer programs to match the terms in the queries with the terms in the descriptions of pages in their databases, and display the titles of the pages in order of estimated relevance to the queries. In the field of information retrieval research, a further distinction is often also made between the internal search logic and processes of the engine and the interface of the search engine – the latter being the means by which the users interact with the former.

The preceding review of the internal and interface components of a search engine above shows that many variables can affect the capability of a search engine to retrieve sources of information on the web in response to a query from a user. Variables include the program logic of the crawler, indexing and retrieval components of the engine, the queries that are permitted by the searcher, etc. Wang et al (1998) reviewed the quality aspects of Internet search engines from the user's point of view. The study highlighted fourteen variables that could be used to assess the quality of a search engine, including the accuracy, reliability, scope of information provided, speed of

response to searches, flexibility and naturalness of the interface, etc. The interplay of these variables means that a search engine might perform very well in retrieving sources for some queries, but also very badly in respect of other queries, and that no search engine will always provide best results. Accordingly, the periodic evaluation analysis of the comparative performance of search engines for different types of queries is of great importancé to users of search engines.

Statement of the Problem

Search engines, meta-search engines and intelligent agents have grown on the Internet in line with explosion in Internet resources and services, including web pages, newsgroups, mailing lists, archives, networked databases, applications, business services, etc. Growth of the Internet and the Web has also led to rapid increases in the quantity and diversity of information sources available on the Internet, including corporate and personal websites, government information, and the online public access catalogues of libraries. Erstwhile publishers of books, journals, magazines and newspapers have also been migrating from print publishing to electronic publishing on the Internet. For users of the Internet, these information sources are at their disposal for doing business, and for searching for information for work, education or leisure.

Although the Internet is considered a universal technology, the extent to which it provides access to electronic information sources from the different regions and cultures of the world varies. English language electronic information resources presently dominate the Internet, and African content in particular is estimated to be extremely low. Adam (n.d), quoting from a July 1998 survey conducted by Network Wizards (a California-based computer firm), reported that Africa was generating only around 0.4 per cent of global content, and a mere 0.02 percent if South Africa is excluded. Jensen (1998) also reported that although the number of African Web sites had been growing rapidly and that almost all countries had local or internationally hosted web servers, the degree of the comprehensiveness of local content on the servers varied greatly. He observed further that there were few well established electronic local content developers and publishers on the continent. In this respect, Chisenga (1998) has noted, for instance, that although African universities and research libraries have over the years collected copies of research reports, theses, dissertations and other documents produced by their students and staff, the web sites of most of the institutions do not provide such information.

The estimated low African content on the Internet might however be due not only to the non-availability of African content on the Internet, but also the inadequate visibility of available African content on the Internet. In other words, African content actually available on the Internet might, for a variety of reasons, be under-represented in the databases of the search engines presently available for searching the Internet. Not much is presently known about the nature and extent of inadequate coverage by search engines of African web sites and/or content. In fact, all the empirical studies that we were able to retrieve from the Internet on the coverage by search engines of region- or country-specific information sources or content had focused on non-African regions and countries.

Accordingly, this study was conceived towards shedding light on the nature of coverage by popular search engines of Africa-specific information content on the Internet. The study sought

to assess and compare the relative effectiveness of major search engines in indexing and retrieving sources of Africa-specific information on subjects or topics of potential interest to African information end-users. There is a further justification for the study. The number of accessible information sources on the Internet is currently astronomical and growing, and different search engines usually give different results for the same queries because they use different strategies for building and retrieving information from their databases. Existing search engines also return long lists of retrieved sources (hits) in response to unsophisticated queries, leaving the average information searcher with a daunting real-time hits evaluation task. As search engines multiply on the Internet, the selection of the most appropriate engine(s) for particular information search requirements becomes a potentially frustrating challenge for the information seeker on the Internet. The problem is particularly serious for the average African information searcher, who is often not very sophisticated in regard to searching for Internet information and might also not have the time and money for a painstaking use of a 'randomly' selected search engine to find the particular information. Such a searcher clearly needs information on the relative effectiveness of different search engines in retrieving Africa-specific information sources on the Internet.

Literature Review

The world is witnessing the development of the global information infrastructure (GII), a computing and telecommunications infrastructure which supports the development, implementation and interoperability of existing and future information services and applications within and across the telecommunications, computing, consumer electronics and content provision industries (ISO JTC, 1996). The Internet is clearly a very important aspect of the evolving GII. Clifford (1998) highlights the role of the Internet as a data transport system, whereas Obenaus (1994) characterizes the Internet as a self-organizing network of networks which, through the interconnectivity it provides between different computer platforms, has attracted the attention of a large number of users. The Internet is now widely used in education, journalism, and research as well as for commerce and entertainment. He emphasizes further that the Internet can however only become a universal information treasure throve if there is active participation on it by all (regions, cultures and social classes).

In spite of the global spread of the Internet infrastructure, it is generally recognized that the Internet is currently dominated by English language content, as well as by content that targets the needs of users in the United States and United Kingdom. The dominance of English language content and search engines on the web is of concern to champions of global cultural diversity. Vaughan and Thelwall (2004), citing Introna and Nissenbaum (2000), note also that differential coverage of web sources by search engines might narrow or bias the universality of the Web by marginalizing certain types of information, for example minority interests or pages in developing countries. It is for this reason that many national governments and international cultural organizations have been promoting initiatives to encourage peoples of different countries and cultures to publish their local content on the Web not only in English language, but also in national and local languages, thereby contributing to the cultural diversity of content on the Web. Jensen (1998) reports, for instance, that French speaking countries in Africa have a higher profile on the Web and greater institutional connectivity than the non-French speaking countries largely due to the strong assistance provided by the various Francophone support agencies, and

the Canadian and French governments, which are concerned about the dominance of English on the Internet.

The preceding observations highlight two very important dimensions of a search engine's performance: (i) its relative coverage of the sources available on the Web, and (ii) its relative coverage of sources in different regions and languages. In respect of the first dimension, Bergman (2001) has pointed out that most of the Web's information is not reachable by standard search engines because the pages may not exist until they are created dynamically by web servers as the result of a specific search. This makes the deep web hidden or invisible. He reports that a study by BrightPlanet Corporation estimated from data collected in March 2000: that public information on the deep Web was then 400 to 550 times larger than the then commonly defined Web; that the deep Web contains nearly 550 billion individual documents compared to the one billion of the surface web; that on the average, deep Web sites receive fifty per cent greater monthly traffic than surface sites; and that the deep Web is the largest growing category of new information on the Internet.

In relation to the second dimension, Lawrence and Giles (1999) found that search engines were more likely to index sites that had more links to them (i.e. the more popular a site the more likely it would be indexed by a search engine). Thelwall (2000) compared search engine coverage of 42 countries and found substantial differences in the coverage of the countries by the five engines tested - AltaVista, Hotbot, InfoSeek, MSN and Yahoo!. Vaughan and Thelwall (2004) investigated the relative coverage by three search engines (Google, AltaVista, and AllTheWeb) of the commercial web sites of four countries (USA, China, Singapore, and Taiwan.) and two languages (English and Chinese). The study concluded that search engines do not cover all of the Web sites available, or even all of the Web sites or pages that they know about from the links in their own databases. The study reported that about 61% of sites in the study were indexed on the average by the three search engines, with Google leading, followed in order by AllTheWeb and AltaVista. The study also found a very strong pattern of uneven coverage of the four countries, with U.S. sites getting much higher coverage than those of China, Taiwan and Singapore. This was true whether the coverage is measured by the percentage of sites covered or the percentage of pages on a site that are indexed. Typically 89% of pages on a U.S. site were covered, whereas only 22% of pages from China and 3% of pages from Taiwan were covered. More than half of the sites from Singapore were not covered at all. The study found no significant relationship between the language of a site and extent of coverage, but found a significant relationship between link counts to a site and the site's coverage by search engines, a finding that is consistent with that of Lawrence and Giles (1999) reviewed above. Vaughan and Thelwall (2004) also noted that their findings were broadly consistent with those of Bharat et al. (2001) and Thelwall (2002) that showed that sites tend to link more within their own country than outside.

Lawrence and Giles (1999) and Vaughan and Thelwall (2004) have observed that differential coverage of Web sources by search engines can have important socio-economic and political implications for different companies, countries and regions of the world in that searchers might be presented with information from only particular sources. Vaughan and Thelwell (2004) note, for instance, that, from a strategic economic perspective, it would be a cause for concern for, say businesses as an example, if users searching for a product online were only pointed to a set of

sites in a particular region because the search engine that they used had not indexed sites in other regions offering the same product. In this regard, Sullivan (2001) has observed that although some of the major search engines have begun to introduce national and linguistic variants of their engines, the variations are often only in the interface alone, with the databases underneath being common to all versions.

Objectives and Scope of the Study

The main objective of this study was to ascertain and compare the extent to which different Internet search engines cover, index and retrieve Internet sources of information on selected Africa-specific subjects. The study therefore sought an answer to the following research question in the context of Africa-based information searchers who may want to use one or more major Internet search engines to find Africa-specific information sources on different highly Africa-pertinent subjects:

> *How do the major Internet search engines compare in terms of their relative abilities to retrieve Africa-specific information sources on different Africa-pertinent subjects?*

Definition of Terms

The study adopted the following definitions for the terms *Africa-based, Africa-specific* and *Africa-pertinent. Africa-based information searcher:* was defined as an information searcher (worker, student, etc) who searches for information for research or decision-making in the African setting. *Africa-specific information* was defined as information on or about Africa or African regions, countries, etc, although such information may not necessarily be found in a web page published by an African organization. *Africa-pertinent subject* was defined as a subject or topic which is often associated with Africa or researched or discussed by Africa-based students, researchers or decision-makers.

Methodology

The study employed the following complementary strategies and tools for data collection and analysis so as to ensure the credibility and reliability for the findings and conclusions.

(i) Survey of the web crawling, indexing and retrieval features of search engines generally:
A comprehensive search for and review of the literature of search engines in general was undertaken. The review considered how search engines find and index sources in their databases, how they search, retrieve and display sources in response to different types of queries, as well as the types of queries that their user interfaces supported.

(ii) Identification and selection of search engines and search terms:
A feeder study was undertaken of the popularity of the use of different search engines, as well as the frequency of search for different subjects or topics, among the postgraduate students of a Nigerian university – the University of Ibadan. The feeder study was designed to answer the following two relevant questions in the context of the present study: (i) Which search engines do the students use most often to search for information on the Internet? (ii) On what subjects or topics do the students seek information on most frequently? The rationale for the feeder study

was to determine the popularity of search engines and search subjects or topics among the students, as representatives of Africa-based searchers for web information sources on Africa-specific subjects or topics. The findings of the feeder study showed that most of the students used either Google (74.1% of the 297 responding students) and/or Yahoo! (22.9%). Other relatively less frequently used engines were Alta Vista (1%), Ask (0.7%) and Lycos (0.7%). [Details of findings of the feeder study are also being published [Salako and Tiamiyu, forthcoming)]. The feeder study provided information on what could be regarded as the most frequently used search engines as well as some of the most frequently searched subjects by the students, and these were adopted as the focal search engines and search subjects in this study.

(iii) Practice, use and review of the selected search engines:
This entailed the practice, use and review of the available literature and online information on the three search engines that were eventually selected for the study. Three criteria were set for the selection of search engines for the study; that the search engine:

(a) must be very popular with the students surveyed in the feeder study (local popularity criterion).
(b) must have been considered a major search engine by experts on search engines (global popularity criterion). We used information from the *Search Engine Watch* (www.searchenginewatch.com) (a highly reliable source of information on Internet search engines) to determine the global popularity of search engines. Sullivan (2004) notes that major search engines of the web are so considered "*because they are either well-known or well-used,* and that webmasters consider the major search engines as the most important places to be listed, because of the traffic they generate, and for searchers, "*well-known, commercially-backed search engines generally mean more dependable results.*"
(c) must have a user interface that enabled *'in title only'* searches. This interface requirement was introduced in order to limit the search engines to matching query terms with words in the titles of sources only during the retrieval tests of the engines. This was in order to prevent the engines from retrieving sources that merely mentioned the query terms in passing in their full text. The assumption was that only the sources that had the query terms in their titles would have contents most likely to be substantially on the subject(s) implied by the terms.

Only *Google* and *Yahoo!* met all the three criteria for selection for study. *Ask* met the global popularity and interface criteria, but not the local popularity criterion, but was however also selected for the study in order to be able to compare retrieval performances between the locally high and less popular search engines.

Search Terms and Queries

A search term is the basic building block of a Boolean or a weighted search. In a search engine, a search term is typically a word, phrase or pattern match expression. The search terms used in the retrieval tests of the search engines were derived through two processes (i) subjects that were searched most frequently on search engines by the sampled students in the feeder study, and (ii) keywords from the subject categories, headings or labels of the directories of different Internet search engines, including those selected for this study.

Subject topics identified in the feeder study were used as reference concepts to browse the subject category labels of different search engines to identify synonymous or near synonymous key words that could have been used to index the subjects by the search engines and/or that could be used by information searchers to retrieve sources on the subject. Eventually, the following twenty key words were selected randomly from the list of identified keywords, and subsequently used to build queries for the search of the engines.

museum	*government*	*health*	*history*	*internet*
education	*poverty*	*politics*	*weather*	*environment*
religion	*corruption*	*human rights*	*culture*	*tourism*
children	*women*	*war*	*HIV*	*refugee*

In building the search queries, each of the above terms (referred to as *subject areas*) was combined (i.e. AND-ed) with the regional terms *Africa*, *Europe*, *(Africa* NOT *south)*, and *Nigeria.* The use of the term *Europe* was in order to compare search results for that region with those for Africa, whereas the use of *(Africa* NOT *south*) was in order to compare results for Africa (inclusive of southern Africa) with those for Africa excluding South Africa. It should also be noted that the terms *Europe*, *Africa*, *south* and *Nigeria* were specified as search terms, and not as geographical restrictions on the queries.

Search Tests

Search tests here refer to the actual information retrieval processes that were performed to test the capabilities of the search engines to retrieve sources of Africa-specific information. The following testing and data collection and comparison procedures were performed:

(i) Test of each of the search engines using queries containing the:
  (a) *subject area* only
  (b) *subject area* AND *Africa*
  (c) *subject area* AND *Europe*
  (d) *subject area* AND (*Africa* NOT *south)*
  (e) *subject area* AND *Nigeria.*

(ii) Recording of the number of sources (hits) returned by each search engine for each search query.

(iii) Comparison of the numbers of hits for *Africa, Europe, Africa* NOT *south* and *Nigeria*, for each subject area and overall, and across the three search engines.

The advanced search facilities of the search engines were used for the tests in that they enabled *'in title only'* searches. Altogether, a total of 100 different queries (20 subject areas, **either alone or AND-ed with each of the f**our regional terms listed in (i) above) were tested on each of the three search engines. The tests were conducted during 12-15 June, 2005. The following are examples of the query syntaxes (using *museum* as the example subject area) used in the searches of each of the three search engines:

Google:

allintitle: museum. This searches the web and returns a list of items representing links to information sources having *museum* as part of the terms in their titles.

allintitle: museum Africa. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles.

allintitle: museum Africa –south. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles but not having *south* in their titles (this attempts to exclude *South Africa*(n) items from the retrieved items).

Yahoo!

intitle: museum. This searches the web and returns a list of items representing links to information sources having *museum* as part of the terms in their titles.

intitle: museum intitle: Africa. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles.

intitle: museum intitle: Africa intitle: -south. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles but not having *south* in their titles (this attempts to exclude *South Africa*(n) items from the retrieved items).

Ask:

intitle: museum. This searches the web and returns a list of items representing links to information sources having *museum* as part of the terms in their titles.

intitle: museum Africa. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles.

intitle: museum Africa –south. This searches the web and returns a list of items representing links to information sources having *museum* and *Africa* as part of the terms in their titles but not having *south* in their titles (this attempts to exclude *South Africa*(n) items from the retrieved items).

Limitations of the Methodology

Firstly, the study focused on comparing the search engines on their abilities to index and retrieve sources on the basis of searches within the titles of the sources in their databases. This *'in title only'* restriction has some undesirable implications, which were however unavoidable given the time-constrained circumstances of the study. One implication is that sources that had titles which were unrepresentative of their contents (and hence, unlikely to have had title words matching the query terms) would have been missed by the search engines despite the sources actually being in their respective databases. Secondly, only the regional terms '*africa*', '*europe*', *'nigeria'* and '*south'* were combined with each *subject area* in the tests, which implied, for example, that a source titled, *'Poverty in the Volta region of Ghana*', would have been missed by a search engine unless the search engine had an internal term mapping mechanism to automatically broaden the search by mapping the narrower country term '*ghana'* into the corresponding broader regional term '*africa*'. We pragmatically opted not to use sub-*Africa*, sub-*Europe* and sub-*Nigeria* entities (e.g. names of specific countries, places or personalities) in the queries because that would have meant having to deal with a virtually infinite number of possible queries.

Test Results

Each table below summarizes the outcome of the retrieval tests in terms of the number of sources retrieved (hits) reported by the three search engines. Additional columns are also provided in

some of the tables to express the *Yahoo!* and *Ask* hits as ratios of the Google hits, thereby facilitating the comparison of the test results across the search engines. Nevertheless, it is very important to bear in mind that the hits for the engines tell us only about the *quantities* of sources retrieved by the engines for similar queries, but not the *quality*, *substance* or *relevance* of the items retrieved.

(1) Subject area only

Table I shows the hits returned by the engines for each of the 20 subject areas. The ratios in the last two columns show that Goggle outperformed Yahoo! substantially in respect of *education* and *museum*, but only slightly so in respect of *environment, health, poverty* and *human rights*. Conversely, Yahoo! outperformed Google on 12 of the 20 subject areas, and strikingly so in respect of *internet, weather* and *women*. Google also outperformed Ask in almost all the subject areas, except *weather*, *religion* and *culture*. The data in the last row of the table provide some indication of the relative performances of the engines on all 20 subject areas, with Yahoo! outperforming Google overall by a factor of 23%, and Goggle in turn outperforming Ask by a factor of about 46% (last two cells of the row).

Table I: Search results - *subject area only*

Subject area	*Google*	*Yahoo!*	*Ask*	*Yahoo!/Google Ratio*	*Ask/Google Ratio*
Museum	10,100,000	5,650,000	2,244,000	0.56	0.22
Government	7,890,000	13,000,000	3,856,000	**1.65**	0.49
Health	32,000,000	27,900,000	11,610,000	0.87	0.36
History	18,700,000	17,600,000	14,610,000	0.94	0.78
Internet	25,200,000	59,900,000	11,550,000	**2.38**	0.46
Education	34,800,000	17,100,000	11,550,000	0.49	0.33
Poverty	696,000	620,000	249,900	0.89	0.36
Politics	3,030,000	4,570,000	2,393,000	**1.51**	0.79
Weather	10,200,000	22,500,000	13,060,000	**2.21**	**1.28**
Environment	4,460,000	3,360,000	1,974,000	0.75	0.44
Religion	1,360,000	2,540,000	1,516,000	**1.87**	**1.11**
Corruption	206,000	246,000	192,400	**1.19**	0.93
Human rights	1,230,000	1,060,000	439,500	0.86	0.36
Culture	4,430,000	4,480,000	4,698,000	**1.01**	**1.06**
Tourism	3,320,000	3,740,000	2,319,000	**1.13**	0.70
Children	7,160,000	7,590,000	3,980,000	**1.06**	0.56
Women	9,550,000	20,500,000	7,553,000	**2.15**	0.79
War	4,450,000	8,280,000	4,464,000	**1.86**	**1.00**
HIV	1,240,000	1,190,000	553,500	0.96	0.45
Refugee	145,000	174,000	92,000	**1.20**	0.63
Overall	180,167,000	222,000,000	98,904,300	**1.23**	0.54

Figures in bold correspond to subjects in which Ask and Yahoo! outperformed Google.

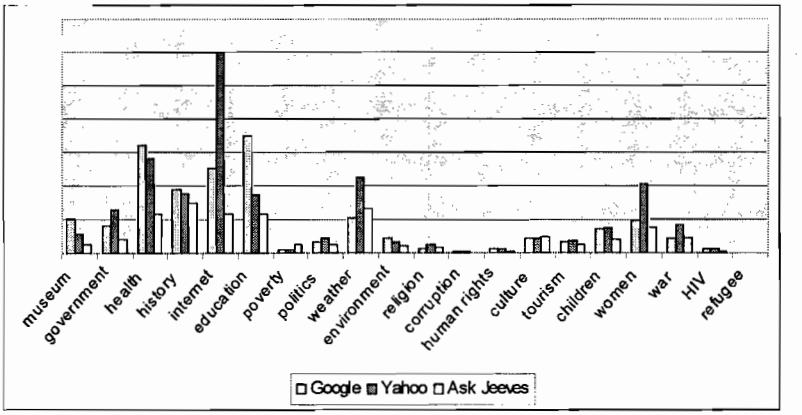

Fig. 1: **Search results** - ***subject area only***

2. Subject area AND Europe

Table II summarizes the data for searches with the query *subject area* AND *Europe*. Analysis of the data shows that Google outperformed Yahoo! in respect of only five subject areas - slightly in respect of *museum, weather* and *human rights*, and substantially in respect of *religion* and *poverty*. Conversely Yahoo! outperformed Google substantially on *Internet, politics, women, tourism, war* and *HIV*. Ask also outperformed Google in 10 of the 20 subject areas, including *tourism, refugee, culture* and *government*. In other words, both Yahoo! and Ask outperformed Google on searches on the subject areas in relation to *Europe*, and by an overall average of about 44% and 35% respectively, as shown in the last row of the table.

Table II: Search results - *subject area AND Europe*

Subject area	*Google*	*Yahoo!*	*Ask*	*Yahoo!/Google ratio*	*Ask/Google Ratio*
Museum	5220	4250	3280	0.81	0.63
Government	26100	45600	55500	**1.75**	**2.13**
Health	37500	59400	11300	**1.58**	0.30
History	39700	50800	81300	**1.28**	**2.05**
Internet	42200	92400	33800	**2.19**	0.80
Education	96600	117000	63600	**1.21**	0.66
Poverty	1060	479	1310	0.45	**1.24**
Politics	13400	28200	22300	**2.10**	**1.66**
Weather	37600	30700	16500	0.82	0.44
Environment	21300	29000	20800	**1.36**	0.98
Religion	65300	14700	27700	0.23	0.42
Corruption	622	825	850	**1.33**	**1.37**
Human rights	5590	4900	6500	0.88	**1.16**
Culture	65200	124000	156600	**1.90**	**2.40**
Tourism	66900	134000	237300	**2.00**	**3.55**
Children	7100	9560	6400	**1.35**	0.90
Women	11900	24100	9730	**2.03**	0.82
War	19000	37800	2870	**1.99**	0.15
HIV	604	1120	920	**1.85**	**1.52**
Refugee	406	492	1360	**1.21**	**3.35**
Overall	563302	809326	759920	**1.44**	**1.35**

Figures in bold correspond to subjects in which Ask and Yahoo! outperformed Google.

3. *Subject area AND Africa*

Table III summarizes the results for searches with the query: *subject area AND africa.* The data show that Yahoo! outperformed Google in respect of virtually all the subject areas, except, and in order, *museum, religion, poverty* and *weather.* However, Google outperformed Ask in respect of 11 of the 20 subject areas, including and in order, *internet, museum, poverty, human rights, tourism* and *HIV.* Focusing on the subject areas that could be regarded as closely related to Africa's developmental problems, the data show that Google outperformed both Yahoo! and Ask on *poverty.* It also outperformed Yahoo! alone on *religion,* and also outperformed Ask alone on *children, women, HIV, education* and *environment.* Overall, with regard to all the 20 subject areas, Yahoo! outperformed Google by a factor of 30%, whereas Ask under performed Google by a factor of about 13% (last row of table).

Table III: Search results - *subject area AND africa*

Subject area	*Google*	*Yahoo!*	*Ask*	*Yahoo!/Google ratio*	*Ask/Google Ratio*
Museum	6010	2400	3060	0.40	0.51
Government	13000	27300	17800	**2.10**	**1.37**
Health	31900	41900	25800	**1.31**	0.81
History	24800	31100	34700	**1.25**	**1.40**
Internet	41400	43300	9500	**1.05**	0.23
Education	23700	31900	21000	**1.35**	0.89
Poverty	9530	7290	5460	0.76	0.57
Politics	9440	13500	9760	1.43	**1.03**
Weather	21500	18500	33600	0.86	**1.56**
Environment	11500	19600	10000	1.70	0.87
Religion	3670	2140	5210	0.58	**1.42**
Corruption	537	1070	848	**1.99**	**1.58**
human rights	11800	15100	7000	**1.28**	0.59
Culture	13200	23700	21700	**1.80**	**1.64**
Tourism	35000	38600	23000	**1.10**	0.66
Children	11400	19100	8860	**1.68**	0.78
Women	14700	22400	10800	**1.52**	0.73
War	10000	17700	10000	**1.77**	**1.00**
HIV	19800	29200	13400	**1.47**	0.68
Refugee	697	758	787	**1.09**	**1.13**
Overall	313584	406558	272285	**1.30**	0.87

Figures in bold correspond to subjects in which Ask and Yahoo! outperformed Google.

4. Subject area AND (Africa NOT south)

These search engines were tested with this query in order to determine the extent to which the search results for the query *subject area AND Africa* in the previous paragraph might have been dominated by hits pertaining to South (or southern) Africa.

The data in Table IV show that Google outperformed both Yahoo! and Ask on *museum, poverty, weather* and *corruption.* Google also substantially outperformed Yahoo! on *government,* but was outperformed by Yahoo! on all other subject areas, including *children, women, war, HIV* and *religion.* Google outperformed Ask in 15 of the 20 subject areas, the exceptions being *government, history, religion, culture* and *tourism.*

Comparison of the data in the last two columns of Tables III and IV show that the Ask ratios were consistently lower for Table IV (*subject area AND (Africa NOT south)),* implying that Ask' coverage of African sources tended to emphasize sources from South (southern) Africa in comparison to Google. Comparatively, Yahoo! to Google hits ratios were higher, lower, or about the same for about equal number of subject areas.

Table IV: Search results - *subject area AND (africa NOT south)*

Subject area	*Google*	*Yahoo!*	*Ask*	*Yahoo!/Google ratio*	*Ask/Google Ratio*
Museum	5030	1750	2040	0.35	0.41
Government	9360	2510	11200	0.27	**1.20**
Health	20900	30600	11200	**1.46**	0.54
History	18600	23100	21200	**1.24**	**1.14**
Internet	13000	21900	3340	**1.68**	0.26
Education	15300	20400	11400	**1.33**	0.75
Poverty	8440	4310	3160	0.51	0.37
Politics	7800	11200	5650	**1.44**	0.72
Weather	4830	2970	2490	0.61	0.52
Environment	9220	10800	6620	**1.17**	0.72
Religion	891	1590	3190	**1.78**	**3.58**
Corruption	694	309	362	0.45	0.52
human rights	10200	13400	3500	**1.31**	0.34
Culture	10300	14700	16300	**1.43**	**1.58**
Tourism	10300	19600	14400	**1.90**	**1.40**
Children	7860	12000	4290	**1.53**	0.55
Women	11600	18800	5890	**1.62**	0.51
War	8020	15300	4780	**1.91**	0.60
HIV	14800	21500	6330	**1.45**	0.43
Refugee	625	738	438	**1.18**	0.70
Overall	187770	247477	137780	**1.32**	0.73

Figures in bold correspond to subjects in which Ask and Yahoo! outperformed Google.

5. Subject area AND Nigeria

Table V shows the results for the searches with the query: *subject area AND Nigeria.* These set of queries were used to narrow down the focus of the search tests to a specific country in Africa (in this case, Nigeria), thereby facilitating a comparison of the results for Africa as a whole with the results for that country. The data show that Google outperformed Yahoo! on as many as 14 of the 20 subject areas, but could only outperform Ask on half of the subject areas, suggesting that Google and Ask were about equally effective in retrieving Nigeria-specific sources. (Notice that the very high overall Ask to Google ratio in the last row of the table is bloated by the abnormally high Ask hits rate for the subject area *weather*). By contrast, Ask to Google ratios were consistently better than the Yahoo! to Google ratios, indicating that Ask performed much better than Yahoo! in retrieving sources on the Nigeria-specific subject areas.

Table V: Search results - *subject area AND Nigeria*

Subject area	*Google*	*Yahoo!*	*Ask*	*Yahoo!/Google ratio*	*Ask/Google Ratio*
Museum	84	19	46	0.23	0.55
Government	3680	1920	1970	0.52	0.54
Health	889	1220	300	**1.37**	0.34
History	559	678	1930	**1.21**	**3.45**
Internet	904	1260	1790	**1.39**	**1.98**
Education	750	1040	1350	**1.39**	**1.80**
Poverty	319	215	371	0.67	**1.16**
Politics	1630	583	1330	0.36	0.82
Weather	668	1740	29400	**2.60**	**44.01**
environment	476	358	406	0.75	0.85
Religion	472	223	401	0.47	0.85
Corruption	515	426	472	0.83	0.92
Human rights	666	1120	1510	**1.68**	**2.27**
Culture	1250	473	1800	0.38	**1.44**
Tourism	3040	334	560	0.11	0.18
Children	834	371	587	0.44	0.70
Women	778	1500	1700	**1.93**	**2.19**
War	755	482	828	0.64	**1.10**
HIV	531	496	894	0.93	**1.68**
Refugee	41	28	37	0.68	0.90
Overall	18841	14486	47682	0.77	**2.53**

Figures in bold correspond to subjects in which Ask and Yahoo! outperformed Google.

6. Comparison of Africa and Europe hits across search engines

Table VI shows the ratios of the hits for the query *subject area AND Africa,* to the hits for the query *subject area AND Europe* in respect of the search engines. The ratios show the extent to which the search engines were able to retrieve African compared to European sources on the different subject areas. Analyses of the data show that all three search engines retrieved, for most of the subject areas, relatively lower African sources than European sources. This was expected as European sources were more likely than African sources to be visible to, and indexed by, the engines. But there were some important departures from this general pattern, particularly in respect of such subject areas as *HIV* (African hits were higher than European hits in the relative order Google, Yahoo!, Ask), *poverty, human rights* and *children* (Yahoo!, Google, Ask), *refugee* (Google, Yahoo!) and *corruption* (Yahoo!).

Table VI: *Africa* to *Europe* hits ratios*

Subject area	Google	Yahoo!	Ask
Museum	**1.15**	0.56	0.93
Government	0.05	0.60	0.32
Health	0.85	0.71	2.28
History	0.62	0.61	0.43
Internet	0.98	0.47	0.28
Education	0.25	0.27	0.33
Poverty	**8.99**	**15.22**	**4.17**
Politics	0.70	0.48	0.44
Weather	0.57	0.60	2.04
Environment	0.54	0.68	0.48
Religion	0.06	0.15	0.19
Corruption	0.86	**1.30**	1.00
human rights	**2.11**	**3.08**	**1.08**
Culture	0.20	0.19	0.14
Tourism	0.52	0.29	0.10
Children	**1.61**	**2.00**	**1.38**
Women	**1.24**	0.93	**1.11**
War	0.53	0.47	**3.48**
HIV	**32.78**	**26.07**	**14.57**
Refugee	**1.72**	**1.54**	0.58
Overall	0.54	0.50	0.36

*Lower ratios imply that retrieved African sources were lower than retrieved European sources, and vice versa. Figures in bold correspond to the few subjects in which Africa-specific hits outnumbered Europe-specific hits.

7. Comparison of Africa and (Africa NOT South) hits across search engines

Table VII shows the ratios of the hits for the query *subject area AND (Africa NOT south)* compared to the hits for the query *subject area AND Africa,* across the search engines. These ratios indicate the extent to which African sources retrieved by the search engines were dominated by sources from South (southern) Africa, with lower ratios implying a lower level of domination, and vice versa. Analyses of the data show that South (or southern) African information tended to dominate retrieved African information retrieved from the three search engines, except for the following subject areas (having ratios of less than 0.40): *internet* (Google only), *weather* (all three engines), *religion* and *tourism* (Google) and *corruption* (Yahoo!).

Table VII: ***(Africa NOT south) to Africa hits ratios****

Subject area	*Google*	*Yahoo!*	*Ask*
Museum	0.84	0.73	0.67
Government	0.72	0.92	0.63
Health	0.66	0.73	0.43
History	0.75	0.74	0.61
Internet	**0.31**	0.51	**0.35**
Education	0.65	0.64	0.54
Poverty	0.89	0.59	0.58
Politics	0.83	0.83	0.58
Weather	**0.22**	**0.16**	**0.07**
Environment	0.80	0.55	0.66
Religion	**0.24**	0.74	0.61
Corruption	0.89	**0.29**	0.43
human rights	0.86	0.89	0.50
Culture	0.78	0.62	0.75
Tourism	**0.29**	0.51	0.63
Children	0.69	0.63	0.48
Women	0.79	0.84	0.55
War	0.80	0.86	0.48
HIV	0.75	0.74	0.47
Refugee	0.90	0.97	0.56
Overall	0.60	0.66	0.51

*Lower ratios imply that lower proportions of retrieved African sources were South (or southern) African sources, and vice versa.

Summary and Discussion of Findings

There is no doubt that the popularity of the Internet has grown in line with the expansion in the availability of different types of Internet resources, and particularly, search engines and other online tools for finding specific resources on the Internet. Nowadays, in most instances, one does not have to know beforehand the Internet address(es) that one wishes to get information from because search engines can assist in getting to the desired web pages. Nevertheless, the success of an Internet search through a search engine still depends on the: (a) existence of the desired information sources on the Internet; (b) ability of the search engine to effectively index/ search the web; and (c) ability of a searcher to effectively use the search engine to retrieve the sources that the engine had indexed.

The findings of this study, which focused on ability of the search engine to effectively index/ search the web, may be summarized as follows: (i) that Yahoo! tended to retrieve more hits on most of the subject areas than either Google or Ask; (ii) that the search engines differed sometimes markedly in their ability to retrieve information on the different subject areas; (iii) that all the three major search engines retrieved more Europe-specific than Africa-specific hits on the majority of the subject areas; and (iv) that South (or southern) Africa-specific information tended to dominate the Africa-specific information retrieved by the three search engines.

Are the above conclusions from the findings of this study conclusive? In view of the limitations of the methodology of this study highlighted earlier, one cannot but admit that one cannot jump to definite conclusions on the overall relative effectiveness of the different search engines. For instance, and contrary to the general conclusions in the preceding paragraph, the data in Table V showed that Ask consistently retrieved more Nigerian sources on each of the subject areas than Yahoo!. Secondly, the methodology of this study rested on the assumption that the effectiveness of search engines can be effectively assessed by running search queries through the titles of the web pages that the engines had indexed in their databases. One should also not lose sight of the very broad subject areas (e.g. education) and geographical concepts (e.g. Africa) that were used in the queries. A further assumption of the analyses and conclusions was that the absolute number of hits returned for identical queries by different search engines could be used to assess the relative coverage of sources by the engines. Actually, the analyses in this study only compared the search engines on their relative abilities to identify, index and retrieve quantities of sources (which emphasized the absolute recall criterion), but not their relative abilities to identify and retrieve qualitative sources and/or display the sources in some order of quality (which would have emphasized such criteria as the relevancy or pertinence of the retrieved sources to the information needs of real-life information end-users. More research is clearly necessary before definitive conclusions can be reached.

Accordingly, the findings of this study can best be regarded as the tip of the iceberg in relation to the infinite possibilities for testing with a greater diversity of general and specialized search engines, as well as with more complex search queries containing narrower subjects and the names of individual African countries, regions, peoples and other entities.

Conclusion

Search engines have become important search tools in the modern era, much in the same way as printed library catalogues, directories and special bibliographies were crucial to finding information in the pre-digital eras. This study was conceived and implemented to investigate, using a few search engines and simple queries, the extent to which the engines are able to index African sources on the Web. As an initial quantitative investigation into the relative coverage of sources of Africa-specific subjects by major search engines, the study investigated only three search engines and twenty broad but non-trivial subject areas. Further research involving more search engines and subject topics, and in respect of sub-continental geographical areas and other entities may be warranted and may lead to different conclusions. It might also be possible to evaluate the quality of some of the sources retrieved by search engines in terms of the relevance of the sources to actual information end-users or in terms of other criteria. These are fertile future research areas due to the increasing importance of search engines and the Web in general.

REFERENCES

Adam, L. (n.d). Giving the Internet an African voice: expanding African content is as important as widening access, **Africa Recovery Online – A United Nations Publication**, Vol. 12 No. 3, available at www.un.org/ecosocdev/geninfo/afrec/vol12no3/internt2.htm (Accessed: 10 October 2005).

Bergman M.K. (2001). The deep Web: surfacing hidden value (Internet), BrightPlanet Corporation, available at www.press.umich.edu/jep/07-01/bergman.html (Accessed: 14 June 2005).

Bharat, K., Chang, B., Henzinger, M., and Ruhl, M. (2001). Who links to whom: mining linkage between web sites (Internet), **IEEE**, available at www.theory.lcs.mit.edu/~ruhl/papers/2001-icdm.pdf (Accessed: 27 June 2005).

Chisenga, J. (1998). A study of university libraries' home pages in Sub-Saharan Africa, ***Libri,*** Vol. 48 No. 1, pp. 49–57.

Clifford, L. (1998). The evolving Internet: application and network service infrastructure, **Journal of the American Society for Information Science,** Vol. 49 No. 11, pp. 961-972.

Introna, L. D. and Nissenbaum, H. (2000). Shaping the Web: why the politics of search engines matters, **The Information Society** (Internet), Vol. 16, pp. 169-185, available at www.scils.rutgers.edu/~belkin/612-05/introna.pdf (Accessed: 27 June 2005).

ISO JTC 1 (1996). **International Standard Organization Joint Technical Committee Report Outline** (Internet), available at www.y12.doe.gov/sgml/wg8/documents/sc18/disk44/18n5580/18n5580.rtf (Accessed: 27 June 2005.

Jensen, Mike (1998). The Africa Internet – a status report (Internet), available at www3.sn.apc.org/africa/afstat.htm (Accessed: 27 June 2005).

Lawrence, S. and Giles, C. L. (1999). Accessibility of information on the Web, *Nature*, (Internet) Vol. 400, pp. 107-109, available at http://pages.stern.nyu.edu/~vassalos/accessibility.pdf (Accessed: 27 June 2005.

Obenaus, G. (1994), The Internet – an electronic treasure trove, **ASLIB Proceedings,** Vol. 46 No. 4, pp. 95-100.

Salako O. and M. A. Tiamiyu (forthcoming). Internet Search Engine Utilization by Postgraduate Students of the University of Ibadan, Nigeria.

Schwartz, C. (1998). Web search engines, **Journal of the American Society for Information Science**, Vol. 49, No. 11, pp. 973-982.

Sullivan, D. (2001). AltaVista Regional Listings Left to Rot, **Search Engine Watch** (Internet), available at http://searchenginewatch.com/sereport/01/09-altavista.html (Accessed: June 27, 2005).

Thelwall, M. (2000). Commercial Web sites: lost in cyberspace? **Internet Research: Electronic Networking and Applications,** Vol. 10 No. 2, pp. 150-159. Also available at www.emeraldinsight.com/Insight/ViewContentServlet?Filename=/published/emeraldfulltextarticle/pdf/1720100205.pdf (Accessed: 4 April 2006)

Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking, **Journal of Documentation**, Vol. 58 NO. 5, pp. 563-574. Also available at www.scit.wlv.ac.uk/~cm1993/papers/2002_Existence_of_geographic_trends_jdoc.pdf (Accessed: 4 April 2006)

Vaughan, L. and Thelwall, M. (2004), Search engine coverage bias: evidence and possible causes, **Information Processing & Management**, Vol. 40 No. 4, pp. 677-692, also available at www.scit.wlv.ac.uk/~cm1993/papers/search_engine_bias_preprint.pdf (Accessed: 23 March 2006)