# Genitourinary medicine and The Internet No 4

R K W Lau

It has been estimated that there are probably in excess of 100 million ($10^8$) Web documents on the Internet, making it virtually impossible for this amount of information to be manually compiled and indexed. In previous articles, I have attempted to catalogue limited lists of sites which may be relevant to genitourinary and HIV medicine. However, there are many occasions where a search for a particular article or documents on a specific subject needs to be conducted. In this article I hope to provide users with some idea of how this enormous volume of data can be selected and retrieved by using a number of facilities freely available on the Internet.

*Internet search engines*

A number of Internet indexing services provide a way for searching and retrieving data. These services provide query forms which drive powerful search engines, consisting of robotic programs which regularly and systematically traverse the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced. They can also be used to search for Usenet articles and documents. Web robots are sometimes referred to as Web Wanderers, Web Crawlers, or Spiders and are principally used for indexing Web sites. They are also useful for validating hypertext links and monitoring new or recent submissions on the Web. In general, robots start a search from a historical list of Web addresses or URLs (Uniform resource locators), especially of documents with many links elsewhere, such as server lists, "What's New" pages, and the most popular sites on the Web. Most indexing services also allow URLs to be submitted manually, which will then be queued, visited and indexed by the robot. Sometimes other sources for URLs are used, such as scanners through Usenet postings and published mailing list archives.

Given those starting points a robot can select URLs to visit and index, and decide to insert it into its database. How this is done depends on how the robot is programmed: some robots index the HTML Titles (HTML, hypertext markup language—the way in which Web documents are written) or the first few paragraphs, or index all works, with weightings depending on HTML constructs, etc. Some use the HTML META tag or other special hidden tags. For a number of reasons, Web server administrators may exclude robotic intrusion. In which case, no information from that site will be indexed although if the exact URL is known, public browsing may still be allowed.

Simple queries are conducted in plain language by submitting a word or phrase to find the relevant documents, although many sites offer advanced search techniques which require a knowledge of Boolean logic. Through a scoring algorithm, the more times a word or phrase appears on a Web page or document, the more likely that word or phrase is to appear at the top of a search for it. Unfortunately, unlike Medline, where bibliographic information such as title, author, publication, date of publication, and keywords can be used to restrict searches, the largely unregulated nature of the Internet and the inconsistent methods by which titles and headings of Web pages are composed by different authors or indexed by Web robots make it more difficult to restrict searches for relevant information, which is occasionally of variable quality. If a query submitted to one search site proves unfruitful, it may be worthwhile trying another.

In view of the sheer size of the Web, the computing resources required to build such large databases and indices of sites and the information they hold, are immense. For example, one of the busiest and largest sites at Alta Vista (`http://www.altavista.com/`) is sponsored by a computer manufacturer and

*Table 1 List of Search Engines on the Internet*

| *Search Engine* | *Web address (URL)* | *Comments* |
|---|---|---|
| Cyber411 | `http://www.cyber411.com` | Performs a parallel search of 15 search engines. Can be very slow at peak times |
| DejaNews | `http://www.dejanews.com` | Dedicated to searching newsgroups |
| Excite | `http://www.excite.com` | A popular site to initiate a search, but can be slow at times |
| Hotbot | `http://www.hotbot.com` | Provides a comprehensive query form to submit a search, and allows users to save their search preferences and session settings |
| Ultraseek | `http://www.infoseek.com/Home?pg = ultra_home.html` | Infoseek's new high-speed "intelligent" search engine provides a good alternative to Alta Vista |
| Yahoo | `http://www.yahoo.com` | Good, moderated catalogues of sites. Limited options for detailed searches |

**Department of Genitourinary Medicine, Beckenham Hospital, Croydon Road, Beckenham BR3 3QL, UK**
R K W Lau

Accepted for publication
10 December 1996

*Table 2 Simple and advanced queries with Alta Vista—rules operating on words, capitalisation and wildcards (*—notation)*

| | | |
|---|---|---|
| Words | Alta Vista treats words as alphanumeric characters separated either by punctuation or white space. Thus *Reiter*, *NSU*, and *KS* are each considered to be single words but *Reiter's*, *N-SU* and *Kaposi's* are all considered to be two words because of the internal punctuation | |
| Capitalisation | Typing a query with upper-case words or phrases forces an *exact* match on the word. To maximise a search, it is generally recommended to type in a word all in lower-case, as this yields case-insensitive matches. Thus, the word *stavudine* in a query should yield matches to *Stavudine, STAVUDINE, stavudine*; or even *sTaVUdinE*; whereas a query with *Stavudine* will only search and list for *Stavudine* and no other match | |
| The * -notation | The asterisk (*)-notation is used to search for similar words or words where there may be variations in spelling. For example, *vagin** will list documents with all possible inflexions, including *vagina, vaginitis, vaginismus, vaginosis, etc.* This may require further conditional restriction (see below) if relevant documents are to be listed. Unfortunately, the * -notation can only be used after at least *three* letters, so it cannot be used to select both *anaerobic* and *anerobic*, whereas a query using *gonorrh* * ea would select both *gonorrhea* and *gonorrhoea* | |
| *Advanced queries only*<br>*Operator* | *Example* | *Finds:* |
| AND | gonorrhoea AND treatment | Pages that include both of the words |
| OR | ks OR kaposi | Pages that include either of the words or both. Note: almost all search engines perform OR searching by default, so it is usually unnecessary to specify an OR search |
| NOT | syphilis NOT yaws | Pages that include the first word (syphilis) but not the second |
| NEAR | mycoplasma NEAR urethritis | Pages in which both words appear within ten words of each other and in either sequence |
| "..." | "bacterial vaginosis" | Pages containing the exact phrase. This is probably the best method of avoiding any ambiguity when searching for a phrase |
| (...) | "genital warts" NOT (cervix OR cancer) | Pages containing the first word or phrase NOT either of the other two. Parentheses simplify the creation of complex queries and can be used in combination with any of the search operators on this list |

consists of six high-capacity servers containing some 30 million indexed pages and handling approximately 20 million queries a day. Its robot, nicknamed Scooter, is reputed to read three million Web pages a day and indexes every word of every page it "reads"! The query examples chosen here are based on Alta Vista—there are other search engines on the Internet, but none match the speed or functionality of its query interface. Table 1 lists a small selection of other Web search engines and their addresses.

The Alta Vista Web site is divided into two separate areas, one for "Simple" (`http://altavista.digital.com/`) and another for "Advanced" (`http://altavista.digital.com/cgi-bin/query?pg=aq`) queries. The latter requires the use of Boolean operators, such as AND, OR and NOT. Table 2 lists some of the rules which operate on the construction of queries. It is not meant to be exhaustive and the examples I have chosen are fairly straightforward. Users may obtain more information by retrieving the Advanced query help document at (`http://altavista.digital.com/cgi-bin/query?pg=ah`).

## *Summary*

The Internet provides a rapid facility for accessing a large amount of information which, once found, can be manipulated in a variety of ways. For example, with the authors' permission, Web pages can be retrieved, converted into text files and edited to produce patient leaflets of sexually-transmitted conditions, prescribing information leaflets, clinical guidelines and protocols, etc. Collections of relevant Web sites placed in bookmarks on a Web browser for students to refer to could also form the basis for self-directed learning, and tutorials and small group teaching on specific subjects.

RL thanks CompuServe UK for providing free access to its Internet gateway.