Andrew J. M. Smith
Emporia State University

УДК 02:004.738.5
https://doi.org/10.18485/climb.2017.5.1.ch1

THE TRUE ART AND SCIENCE OF THE LIBRARY IN THE AGE OF THE INTERNET SEARCH ENGINE

Summary

Libraries are increasingly under pressure to "modernize" and to move away from a book-centred model to a patron-centred model, focusing on the speedy provision of information. The popular perception of the library is that it has been superseded by the internet, and that all information is now freely available online. Think of the internet as a library and the search engine as the catalogue. While much information is indeed freely available online, what is available is still only a small portion of what exists. In addition, the profligacy and disorganization of online information create a substantial barrier to the effective retrieval and use of the desired information, a barrier that is only compounded by the structure and operation of internet search engines.

Deep web

The actual amount of information freely available online is a matter of some debate. In addition to what is termed the visible web, that is the information that is available to automated harvesting and "indexing" by search engines, there is also what is termed the deep, which refers to information that is freely available online, but is not indexed. (Note this is also referred to as the dark web, but in addition to being used synonymously with the deep web, the term dark web is also used to denote a subsection of the deep web where illegal activity takes place.) This information is not static, being generated from databases or not sitting on an easily discoverable place or in a machine readable format, so is not detectable by the spiders, crawlers, or automatic "indexing" bots employed by the major search engines. Nobody appears to be quite sure how large

the deep web is. Everyone agrees that the initial estimates from the turn of the 21st century of between 400 and 500 times the size of the visible web are not correct, but whether they over-report or under-report is not clear.

Estimates of the size of the indexed web as of May 2018, give a figure of at least 1.8 billion pages (de Kunder, 2018), but it is also estimated that Google has to date only indexed anywhere from 0.04% to 5% of the web (Bruce, 2010). This figure does not include web pages with access restricted by pay-walls, by authentication protocols, or other access controls. Again, exact figures are hard to discover, and with the amount of new information being generated daily on the web, the huge numbers of pages or amounts of data discussed can hide the fact that large amounts of data are not available through the web, or information is available, but not available through search engines.

Search Engines – indexes

Any consideration of the usefulness of the internet as a replacement for the library must begin with an understanding of how search engines retrieve data. The process begins with automated search and retrieval of webpage material by software processes called crawlers, spiders or bots. Once material is discovered and retrieved, it is "indexed" and then stored in massive data banks for future matching and retrieval. We have already seen that automated information discovered is limited in scope by the nature of the webpage structure or the type of data presented. The "indexing" of the material discovered is a purely automated process, and consists, according to Google's website, of listing every word that appears in a document. (Google, n.d.)

However, this type of machine-created index results in an product that is substantially different from a human-created one. Essentially a search taking place with a machine-created index is using a full-text matching strategy, where the words in a search query are matched against the words in the "index" and all pages matching the search terms are returned.

Lamb (2004) offers the following five limitations of full-text searching compared to a high-quality human-generated index: Full text searching cannot easily deal with homographs (words that are spelled the same but have different meanings); it cannot accommodate synonyms or, for

example, words in different languages referring to the same geographical location; it does not ascribe value to word occurrences, and so cannot distinguish between important or unimportant references to the search terms; it cannot understand where a topic is inferred but particular search terms are not used; and it cannot index pictures or diagrams – it may be able to search picture captions but not the content of the picture or diagram.

Again, as Lamb (2004) notes, computer produced indexes are more oriented toward information providers, as they can be produced rapidly and inexpensively, and are particularly useful when the content itself is in flux. On the other hand, information users are better served by human-produced indexes, which take longer to produce and are obviously more expensive because of the time spent by professionally-qualified indexers. However, these indexes provide better and quicker access to relevant information although they are by their nature more suited to more stable information collections, or those that are additive in nature.

The work of the professional indexer is much more complex than simply compiling lists of words. As the Society of Indexers of the United Kingdom note,

> An indexer considers the terms the readers are likely to use and relates them to the language chosen by the author. An indexer analyses the meaning and significance of the entire content in detail, and identifies tangible concepts from the woolliest of descriptions. (Society of Indexers, n.d.)

Search engines – how do they search? What do they search

Even for those search engines that provide more than a simple match to words in indexed documents, there are problems in information retrieval. The search algorithms can be extremely complex, constantly changing, and proprietary, so not open to examination. Google, for example, claims that it poses 200 questions in the course of an internet search, and gives a few examples. However, it also states that these are being changed on a regular basis, so the consistency of the search is not ensured. Neither are the search algorithms defined, so the user has no idea which particular criteria are being used to filter results. (Google, n.d.)

The reference interview in itself has always been the means whereby a trained librarian helped the information seeker refine his or her ques-

tion to the point where the information sought was exactly what was required: only relevant information was returned and irrelevant information excluded. It also acknowledged the fact that information seekers often ask overly general questions that do not directly identify the information required. While in some ways the Google approach of 200 questions mimics the reference interview, it cannot elicit any responses from the information seeker and is based therefore on assumptions about the search terms or information seeker and will inevitably result in poorer definition of the desired search with a parallel limiting of the delivered results.

Paid placement, website ranking and searcher profiling

Another complication in the use of search engines from the information-seeker's point of view, is the rise in the use of the internet to generate income from the supply of information, including information that is freely available elsewhere. This can have both positive and negative effects on the access to and provision of high quality information. Monetary considerations increasingly influence the selection and placement of search results in online search engines and raise questions of accuracy, suitability, and the neutrality of the information provider. Some search engines evidently do accept payment for higher placement in results listings, while others do not accept direct payment, but use multiple advertisement placements in and around search results to generate income.

Searcher profiling can also be in use, whereby a person's viewing or searching habits are tracked and only websites with similar viewpoints or content will be returned in searches, thus limiting the searcher's ability to see anything that represents differing or conflicting viewpoints.

Authority control versus social tagging

Access problems are also highlighted in the rapid acceptance of social tagging or folksonomies as a replacement for controlled-vocabulary cataloging or indexing. The use of controlled vocabulary was designed to ensure the retrieval of relevant information and the non-retrieval of ir-

relevant information by the application of agreed-upon subject terms by competent indexers and catalogers. We may argue about the shortcomings of the actual subject terms used for issues such as exclusion, lack of diverse points of view, or cultural insensitivity, as well as the rate of change of the controlled vocabulary thesaurus, but this method provides effective and efficient information retrieval. Social tagging does offer benefits, being created by information users using natural language, and enjoying the ability to be up to date and to change as needed. However, social tagging is not free from the criticisms leveled at the use of controlled-vocabulary indexing, although some of the problems may manifest themselves in slightly different ways. User-generated tags are often overly general, resulting in over-retrieval of unconnected information; overly specific thus limiting retrieval; use new or trending terms for which there is no accepted definition or multiple, conflicting definitions; or using terms or definitions in non-standard ways, resulting in inaccurate retrieval (du Preez, 2015).

Internet as library and search engine as catalog

The comparison of the internet as library with the search engine as catalog misses many of the problems inherent in a system that ignores decades if not centuries of information organization wisdom and we are still seeing retrieval systems that are focused on the number of information items they can retrieve, rather than on the quality or suitability of the items retrieved in response to a particular search. Few would be satisfied with a library catalogue that contained only 5% of the library's holdings or materials accessible through the library, yet we appear to be perfectly content with this standard on the internet. Again, we would be suspicious of any cataloging or retrieval system in the library that was not transparent, where we could see exactly what criteria were being used to classify information and what criteria were used to select information in response to a query. All search engines currently operate "behind closed doors", so the information seeker has no idea what was searched, what was retrieved, what was provided and, more importantly, what was withheld from the results.

The sheer size and variety of the internet hides these facts from the information seeker. A single-word search on the term folksonomies in May

2018 resulted in 2,300,000 hits (Google), 707,000 hits (Yahoo), and 282,000 (Bing). This illustrates both the differences in different search engine results, as well as the overwhelming scale of the number of results being returned in a search. Who could imagine that few or none of the 2 million hits generated by a search engine may give a complete answer to the question posed? How difficult is it, in this age of more is better, to understand that one properly-indexed result of a targeted search in a library database by a trained information professional can provide authoritative information more quickly than the 2 million word-matching hits of a search engine? The internet is an astonishingly powerful tool, but like all tools, it can do better in the hands of a master craftsman than in the hands of a novice.

The current situation highlights the necessity of information seekers having highly developed information literacy skills, as they must navigate through an increasingly dense jungle of information sources that are often poorly organized, include more incorrect or misleading information or where relevant information is hard to retrieve. The challenge for libraries is to educate users to develop these essential information literacy skills, and to promote a deeper understanding of the capacity and the limitations of the internet as an information resource in contrast to the technology-enabled resources available through the library.

The true art and science of the library is revealed in the libraries that are able to harness the power of information technology to streamline information retrieval while at the same time not discarding effective organizational schemes or losing the serendipitous discovery of related information.

References

Bruce, J. (2010, December 31). *18 fun interesting facts you never knew about the internet.* Retrieved from http://www.makeuseof.com/tag/18-fun-interesting-facts-knew-internet/

de Kunder, M. (n.d.) *The size of the World Wide Web (the Internet).* Retrieved May 5, 2018 from http://www.worldwidewebsize.com/

du Preez, M. (2015). Taxonomies, folksonomies, ontologies: What are they and how do they support information retrieval? *The Indexer, 33*(1), 29-37.

Google (n.d.). *How search organizes information*. Retrieved May 10, 2017 from https://www.google.com/search/howsearchworks/crawling-indexing/

Internet Live Stats (n.d.). *Total number of websites*. Retrieved May 5, 2018 from http://www.internetlivestats.com/total-number-of-websites/

Lamb, J. A. (2004, February 10). *What is wrong with full-text searches?* Retrieved from https://jalamb.com/2004/02/10/full_text_searches/

Society of Indexers (n.d.). *Getting from A to Z can be this easy.* Retrieved May 6, 2018 from https://www.indexers.org.uk/wp-content/uploads/2017/03/SI-brochure-no-graphics.pdf

Thomas, R. (2017, May 23). *How do search engines rank websites?* Retrieved from https://www.sourcelinemedia.com/how-do-search-engines-rank-websites/

Ендрју Џ. М. Смит
Државни универзитет Емпориа

ПРАВА УМЕТНИЧКА И НАУЧНА ПРИРОДА БИБЛИОТЕКЕ У ДОБА ИНТЕРНЕТ ПРЕТРАЖИВАЧА

Сажетак

Библиотеке су под растућим притиском да се „модернизују“ и да се удаље од модела са фокусом на књигу ка моделу са фокусом на корисника, концентришући се притом на брзо пружање информација. Популарна перцепција библиотеке јесте да је замењена интернетом, те да су данас информације слободно доступне на мрежи. Премда велики број информација заиста јесте слободно доступан на мрежи, сама природа њихове позамашности и дезороганизације представља значајну препреку за ефикасно проналажење и коришћење жељених информација.

Сведоци смо пораста коришћења интернета у сврху остваривања профита на основу прибављања информација, укључујући информације које су слободно доступне и на другим местима. Ово може имати и позитивне и негативне последице на приступ и пружање квалитетних информација. Финансијски разлози све више утичу на избор и постављање резултата претраге у претраживаче на мрежи и стога покрећу питања о тачности, подобности и неутралности онога ко пружа информације.

Поређењем интернета са библиотеком која поседује претраживач уместо каталога, игноришу се многи проблеми који су својствени систему који не узима у обзир деценијску, ако не и вишевековну традицију мудре организације информација, те смо и даље сведоци система који су фокусирани на количину информација до којих се може доћи, а не на квалитет или подобност информација до којих смо дошли одређеном претрагом. Проблеми приступа информацијама су такође наглашени у брзом прихватању социјалног означавања као замене за контролисану каталогизацију појмова, при чему су ознаке често исувише опште, што доводи до проналажења прекомерног броја неповезаних информација, које су исувише специфичне чиме се

ограничава процес проналажења информација, или употребом термина и дефиниција на нестандардан начин, што за резултат има непрецизно проналажење информација.

Тренутна ситуација указује на чињеницу да особе које траже одређене информације морају имати изузетно развијене вештине информационе писмености, јер су принуђене да пролазе кроз све гушће џунгле извора информација, а које су све лошије организовне, укључују све већи број нетачних и погрешних информација или информација које су тешко доступне.

Права уметничка и научна природа библиотеке се огледа на примеру библиотека које су у стању да искористе моћ информационих технологија у циљу поједностављивања процеса проналажења информација, а које истовремено не укидају могућност случајног откривања сродних информација и не одбацују ефикасне организационе шеме.

Кључне речи: организација знања, претраживачи, претрага, проналажење информација, информациона писменост, случајно откривање.