

Provisional Title: Networked data repository efficiency and effectiveness - A users view.

Information overload is (and always has been) a major issue for anyone searching for a specific piece of information in an ocean of data. The best metaphor is probably 'looking for a needle in a haystack'. And the biggest issue for the user is the 168 hours in each week, i.e. time.

All efficient information retrieval systems rely on understanding what needs to be retrieved - filing systems are about retrieval not storage. A general guide for the efficiency of a filing system is that if the data sought cannot be retrieved, or proved not to exist in the system, in less than 5 minutes, the system is ineffective.

Despite the power of 'search engines', the internet is a repository of both useful, but almost entirely useless information (to any specific individual with any specific query); it was created with transmission and storage, not retrieval, in mind.

Lets use a simple example (and test it if you like): A child finds a beetle that neither they nor their parents recognise, and they want to identify it. They have various possible 'files' they could search, starting with memory and books. They could take it to the local museum in the hope of finding an expert or appropriate repository of specimens for comparison; They could go to the local library, in the hope that one of the texts will have an identification chart (preferably visual, since they are not expert coleopterists (and do not even know this word) and so serious scientific documentation is of little value); They could search online, but then what is the most appropriate search term - the most likely is 'beetle, identification' or similar and that does not lead (in the first 5-6 pages to the 'Coleopterist Society's' page and their 766 photo gallery of beetles); They could ask a friend via social media or in person, but that relies on serendipity.

Theoretically, the internet should be the most efficient if the technologists claim of the usefulness of the web is to be believed. But that is, increasingly, not the case as the net becomes clogged with billions upon billions of interlinked bits of data (most of it useless for any purpose, let alone a specific purpose), that is growing exponentially, and even Google struggles to keep pace with the growth and still provide the best linkages to sites that might answer the child's apparently simple, question of 'what is this bug?'. Furthermore, as the internet matures, access to such data is increasingly becoming 'owned' by a media corporation or a government department that needs to charge a fee (to somebody) to either remain in business or fund their activities.

The issue is confused further by organisations and technologists seeking to understand the way search engines work in order to get their particular site in front of a potential customer (hence the actually useful source of the Coleopterist Society is in competition with commercial interests and falls down the SEO rankings), and so search engines begin to favour the ones in the know, e.g wikipedia (and even that supposedly organised repository cannot address the 'whats this bug' question). And of course it is possible that the data does not exist, but the online searcher has know way of knowing that and the search engine has no way of understanding the exact nature of the query and can only provide possible options not a definitively negative response. Add the rise of social networks and the pressures on time increase.

So, a few central problems emerge. For any organisation seeking to get a message through the fog that obscures the direct and simple transmission of a key message to a specific user desirous of receiving that message, the problem seems almost insurmountable; and for the user seeking to locate a specific piece of information, the search is becoming increasingly costly (in terms of time or money).

Question: Is there an 'efficient' as well as an 'effective' sieving mechanism for retrieval of the specific from the non-specific? Is it about the SE categorisation and prioritisation or is it about the users search skills? What does the user view as an 'efficient and effective' search? etc..

I have some ideas forming around a couple of theoretical perspectives, but seek assistance from someone with a strong interest in search engines, networks, or similar to explore and develop the ideas and to conduct some preliminary research.

David Arnott, WBS