RESOURCE ARTICLE

The chemical safety gateway: Beyond Google's limitations

The dominance of Google as the default search engine (more than 80% of Internet searches) is not based on its ability to deliver technically useful information. Its search strategy, ranking criteria and business considerations, can mask useful chemical safety information from the safety professional, chemical users and the general public. Fortunately, alternative search engines are available to use to collect both raw health and safety data as well as technical information and policy statements to inform risk assessments. It is important for the health and safety professional to be aware of these alternatives to perform efficient searches and to be able to critically assess information collected from the Internet by the workers they serve. This article describes the challenges presented by Google's strategy and an alternative strategy that allows for more strategic search for chemical safety information.

By Abe Lederman, Sol Lederman

INTRODUCTION

Google excels in certain types of web searches and performs poorly in others. Google is efficient and precise in telling us where nearby restaurants are, what their hours are, and in sharing their menus. But, efficiently delivering quality scientific information from chemical safety databases is not an area in which Google excels for a number of reasons. In this paper we cover four topics:

1. Identify deficiencies in Google which hinder research and, worse, cause us to miss relevant information.
2. Introduce distributed search, also known as federated search, as an alternative to Google's "web crawler" paradigm and explain why this alternative is best suited for chemical safety and other scientific and technical research.
3. Present the Chemical Safety Gateway as a novel alternative to Google.
4. Conclude with a discussion of possible next steps for the gateway.

Abe Lederman *is affiliated with Deep Web Technologies, Santa Fe, New Mexico, United States (e-mail: abe@deepwebtech.com).*

Sol Lederman *is affiliated with Deep Web Technologies, Santa Fe, New Mexico, United States (e-mail: sol@deepwebtech.com).*

GOOGLE'S SEARCH DEFICIENCIES

We highlight a number of Google's major search deficiencies then recommend a better strategy to finding quality chemical safety information.

- **Biased ranking.** Google's ranking algorithm was designed to favor search results with the greatest number of incoming links. Thus, results with more links from other sites, especially popular sites, were themselves deemed more popular and they were displayed higher up in the search results. Today, Google favors paid content, i.e., advertising content that generates revenue. The researcher is now required to wade through ads for chemicals, supplies, and safety equipment hoping to find quality, authoritative safety material.
- **The filter bubble.** Google "personalizes" its search results. Wikipedia explains that a filter bubble[1] is "a state of intellectual isolation that can result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user, such as location, past click-behavior and search history." Filter bubbles lead to different sets of results being served to different users, leaving the researcher wondering what content Google believes is relevant to them and what content Google is omitting (or burying) because it deems it less important.
- **Who decides what is important?** Google has amassed information about a huge amount of web content. Because none of the information sources for which Google displays results are curated, much of what Google presents is low quality. Researchers need comprehensive and reliable information sources. Having to sift through the firehose of Google results and judge each result is tiring.
- **Google's crawlers miss much of the high quality content.** This is perhaps the greatest limitation of Google when performing scientific research. Google amasses its huge collection of web content data by following links from one web page to another. But, a large percent of chemical safety information lives in databases behind web forms. Google is not designed to fill out and submit web forms so it misses all of that information. Additionally, much of the highest quality chemical information is only available through a paid subscription. Google will not access that content either.
- **Google presents its results completely unorganized.** Focused and efficient research requires the ability to group search results by

1871-5532
https://doi.org/10.1016/j.jchas.2017.11.001

Please cite this article in press as: Lederman, A., and Lederman, S., The chemical safety gateway: Beyond Google's limitations. J. Chem. Health Safety (2017), https://doi.org/10.1016/j.jchas.2017.11.001

topic, category, source, document type, and other characteristics. Being able to refine results is also important and it relies on grouping of results, especially by topic. Google has no notion of organizing results.

We have highlighted four deficiencies of Google which cause the search engine to either not return relevant results at all or to return them in an unexpected order, potentially mixed with irrelevant results, perhaps buried among ads, and of unknown quality. We recommend "Relying on Google for Science Information is Bad for Your Health"[2] to learn more about Google's deficiencies. "Quality, Not Quantity: The danger of overlooking quality of search results"[3] further discusses limitations of Google, Bing, and other popular search engines and introduces distributed search of deep web databases as a better strategy for finding quality chemical safety information.

DISTRIBUTED SEARCH AS AN ALTERNATIVE TO GOOGLE

Distributed search technology aggregates information from a portion of the web by searching curated content databases and not by arbitrary following links between web pages. New Mexico-based software company Deep Web Technologies [DWT][4] develops distributed search tools that provide an alternative to Google. DWT, in collaboration with Grace Baysinger,[5] Chemistry and Chemical Engineering Librarian at Stanford University, developed Chemical Safety Gateway, a portal which brings together high quality chemical safety content from public and subscription sites. DWT developed a second version of the gateway for the public.[6] The public version searches 46 chemical safety-related information sources. Both sites demonstrate the strengths of distributed search while bypassing the limitations of Google.

In its use of distributed search, Chemical Safety Gateway is fundamentally different from Google in a number of ways:

- **No crawling.** Distributed search engines do not follow links to find documents. They fill out search forms to access content from web databases, in much the same way that humans do.
- **Real-time parallel searching.** Chemical Safety Gateway searches its 46 sources simultaneously. And, because the parallel searches do not rely on Google's index which may have dated content, results are as current as the content in the databases they search. Thus, as soon as database owners update their content, the very next search will retrieve the latest version.
- **Curated sources leads to better results.** Chemical safety experts selected the 46 sources that comprise Chemical Safety Gateway. Searching only curated sources eliminates the Google problem of returning too many irrelevant results from unknown sources and needing to sort through them.
- **Better relevance ranking.** Chemical Safety Gateway ranks search results based purely on their relevance to the user's search terms. Biased ranking based on document popularity and filter bubbles are avoided as is the problem of irrelevant results ranking highly due to commercial interests of the search engine provider.
- **Organized results.** The gateway organizes results by topic, category, source, document type, and other characteristics making it easy to compare results from the same or different sources, to examine documents of different types, to refine results based on topic and to intelligently assess the quality of the information being provided.

Beyond the limitations of Google, distributed search is particularly important because relying too heavily on a single source is problematic. OSHA makes the case, in "Hazard Communication: A Review of the Science Underpinning the Art of Communication for Health and Safety,"[7] that a popular resource, the MSDS, (a term that we use interchangeably with SDS,) is not dependable as a single source of information.

> "MSDSs, by themselves, are a poor means of informing workers of hazards to which they may be exposed for the following reasons: 1) much of the technical information has little meaning to the average worker and may even frustrate the workers' ability to read other portions of the MSDSs that have information pertinent to hazard recognition and safe practices; 2) information depicting hazardous conditions, signs or symptoms of exposure, and safe handling procedures are written generically and workers may have difficulty seeing the connection between their own use of the chemical and the information on the MSDS; and 3) the information may contain terms too difficult to understand or too brief and vague to actually generate the concern that worker should have regarding safe use of the chemicals."

Additionally, the OSHA report found that literate employees understand less than 60% of the content in an MSDS and that only 11% of MSDSs were accurate in all of these areas: health effects, first aid, personal protective equipment and exposure limits.

Unifying results from a number of different sources, each with its particular focus, is critical in painting a clear, complete, and accurate picture of the information being sought.

THE CHEMICAL SAFETY GATEWAY

DWT and Stanford University initiated the Chemical Safety Gateway project in the fall of 2015. The initial deployment combined the public sources with thirty commercial databases to which Stanford subscribes. DWT founder Abe Lederman described the joint work in a presentation, "Unique One-Stop Access to a Multitude of Chemical Safety Resources",[8] which he delivered at a session of the Division of Chemical Health & Safety April 2017 National ACS Meeting.

Chemical Safety Gateway searches sources in seven categories. Table 1 lists representative sources in each of the categories. Sources marked with an asterisk (*) are subscription-only sources and are not available in the public gateway. Additionally, some of the publicly searchable sources require a subscription or pay-per-view to access the content.

Figures 1 and 2 show results from the Stanford Chemical Safety Gateway. Figure 1 shows all results from

Please cite this article in press as: Lederman, A., and Lederman, S., The chemical safety gateway: Beyond Google's limitations. J. Chem. Health Safety (2017), https://doi.org/10.1016/j.jchas.2017.11.001