ETDE-OA—237

ETDEWEB versus the World-Wide-Web: A Specific Database/Web Comparison

Debbie Cutler
Energy Technology Data Exchange (ETDE) Operating Agent
DOE/OSTI, Oak Ridge, TN (United States)

Publication Date: June 2010

Abstract:

A study was performed comparing user search results from the specialized scientific database on energy-related information, *ETDEWEB*, with search results from the internet search engines Google and Google Scholar. The primary objective of the study was to determine if *ETDEWEB* (the Energy Technology Data Exchange – World Energy Base) continues to bring the user search results that are not being found by Google and Google Scholar. As a multilateral information exchange initiative, ETDE's member countries and partners contribute cost- and task-sharing resources to build the largest database of energy-related information in the world. As of early 2010, the *ETDEWEB* database has 4.3 million citations to world-wide energy literature. One of *ETDEWEB*'s strengths is its focused scientific content and direct access to full text for its grey literature (over 300,000 documents in PDF available for viewing from the ETDE site and over a million additional links to where the documents can be found at research organizations and major publishers globally). Google and Google Scholar are well-known for the wide breadth of the information they search, with Google bringing in news, factual and opinion-related information, and Google Scholar also emphasizing scientific content across many disciplines. The analysis compared the results of 15 energy-related queries performed on all three systems using identical words/phrases. A variety of subjects was chosen, although the topics were mostly in renewable energy areas due to broad international interest. Over 40,000 search result records from the three sources were evaluated. The study concluded that *ETDEWEB* is a significant resource to energy experts for discovering relevant energy information. For the 15 topics in this study, *ETDEWEB* was shown to bring the user unique results not shown by Google or Google Scholar 86.7% of the time. Much was learned from the study beyond just metric comparisons. Observations about the strengths of each system and factors impacting the search results are also shared along with background information and summary tables of the results. If a user knows a very specific title of a document, all three systems are helpful in finding the user a source for the document. But if the user is looking to discover relevant documents on a specific topic, each of the three systems will bring back a considerable volume of data, but quite different in focus. Google is certainly a highly-used and valuable tool to find significant 'non-specialist' information, and Google Scholar does help the user focus on scientific disciplines. But if a user**'**s interest is scientific and energy-specific, *ETDEWEB* continues to hold a strong position in the energy research, technology and development (RTD) information field and adds considerable value in knowledge discovery.

Acknowledgements:
The author appreciates the feedback gained from ETDE member organizations throughout the review process of the study, from initial results through publication of the report.

ETDEWEB versus the World-Wide-Web: A Specific Database/Web Comparison

The world-wide-web has dramatically changed the way most businesses and individuals search for information. The web has become an integral part of the daily lives of an ever-growing percentage of the world's population both in the workplace and at home. Search engines such as Google consistently enhance the perception that 'everything is out there' for the taking and free to look at, reporting numerous results for almost any search. Business managers scrutinize library and information resource budgets more carefully than ever, often leaving workers on their own, thinking they can easily find everything they need. Quantity is no problem when it comes to search results; but what about other aspects? Who controls the quality and completeness of information and what search engines 'find'? Who is responsible for getting information out on the web? How much time do individuals now spend individually filtering through this virtual sea of information? Do they believe what they see is the best or most reliable information available? Are specialty databases and information portals more helpful to users? And if so, do they add value?

These questions are of particular interest in the scientific community where quality information sharing is vital to the advancement of science. For the world's largest scientific database in the energy field called *ETDEWEB*[1], a comparative study has been carried out, looking for answers to some of these questions. In the study, *ETDEWEB* search results are compared with results from two of the most popular search engines, Google and Google Scholar (GS). The primary goal was to determine if *ETDEWEB* continues to bring the user search results that are not being found by those search engines, thus illustrating its enduring value. The study concluded that even with the plethora of information on the world-wide-web, *ETDEWEB* continues to be a significant, niche resource to energy experts for discovering relevant energy information. The analysis showed that *ETDEWEB* provides search results that are, on average, unique 86.7% of the time when compared to results for the same search being returned from Google or GS. Much was learned from the study beyond just metrics. Observations about the strengths of each system and factors influencing the search results are also shared. The paper provides background information on the study and a summary of the results.

Study background

The idea for the *ETDEWEB* study was prompted by earlier 2009 analysis comparing search results from WorldWideScience (WWS) (an information portal) to Google and GS results. As a newer information system, WWS sponsors were interested in learning how its contents compared to what could be found on the web. As *ETDEWEB*'s sponsors have also been keen to know how the ETDE database compares to Google/GS, a separate study following a similar methodology was initiated.

ETDEWEB's sponsor is ETDE, the Energy Technology Data Exchange, a multilateral information exchange agreement formed in 1987 under the framework of the International Energy Agency (IEA). ETDE member countries contribute funding, and both members and partners contribute database records/documents representing worldwide energy research, science and technology R&D results, including policy, environmental and economic aspects. ETDE's mission is to provide governments, industry, and the research community in the member countries with access to the information collected and to increase dissemination to developing countries. Over 110 countries have free access to *ETDEWEB*, and ETDE welcomes interest from countries that do not already have access. The ETDE mission is achieved through the collaborative creation of the Energy Database. The web version,

[1] *ETDEWEB* stands for the Energy Technology Data Exchange – World Energy Base

ETDEWEB (ETDE World Energy Base) (http://www.etde.org/etdeweb) was used for the comparison. The study was performed by ETDE's Operating Agent, the Department of Energy's Office of Scientific and Technical Information (OSTI) who maintains *ETDEWEB*.

Understanding the search systems

ETDEWEB is renowned as the single largest database of energy-related information in the world. As of early 2010, the database has over 4.3 million citations to world-wide energy literature. One of *ETDEWEB*'s strengths is its focused scientific content, with the full text for its grey literature (over 300,000 documents in PDF and growing daily) available for viewing directly from the ETDE website. Over a million additional citations contain links directing users to where the documents can be found at research organizations and major publishers globally. *ETDEWEB*'s underlying database structure allows simple and advanced search options and other interface features that aid users in narrowing their results. The information contained in *ETDEWEB* is expected to be of a scientific nature and is filtered by member countries for inclusion. Sources include reports from major research organizations, technical conferences, peer-reviewed journals and much more. A user searching *ETDEWEB* does not have to sift through press releases, product promotions, advertisements, and social media sites to see research results aimed primarily at the scientific community. Documents that are in other languages generally have English titles and abstracts included in the database to help users find relevant subject content and to aid decisions on whether translation of the full text is worthwhile. Subject indexing using a controlled vocabulary is added to each database citation to also help in more precise retrieval. A new subject clustering search tool available in *ETDEWEB* in February 2010 helps narrow searches, taking good advantage of this subject indexing.

Google is a remarkable product that has transformed the user experience, even making its way into common usage as a verb. Its strength lies in the vast amount of information accessed in mere milliseconds and its ranking algorithms that are often uncanny in their ability to deliver information matching users' interests. A typical search generally returns millions of 'hits,' with the first 10 displayed to the user, and the rest available in increments of 10. Google truly does try to index 'everything' it can access. The types of information found in Google include some scientific content from the 'surface' web, but there appears to be a greater focus on news, business information and products, blogs, promotional materials, reviews, information/references sites like Wikipedia, and large sales outlets such as Amazon. Google Scholar, as the name implies, typically focuses on resources targeted to be of interest to the academic and research community. Key journal publishers, information societies and many scientific databases formerly considered part of the 'deep web' are made more readily accessible through this specialty search engine.

The framework for each search system is also useful to understand. Search engines like Google/Google Scholar (GS) are generally not the underlying source of the information content, although they have many partners. The search engines have 'bots' or 'crawlers' that seek out websites and build indexes that help rank and recognize similar items, with the ranking contributing/controlling what the user sees as search results. For much of the data on the web, a person or an entity has to make a deliberate and conscious effort to ensure that the information they produce is put in an acceptable format and is visible to the bots, letting the search engines know where and what to index. Specialty companies claiming to be able to increase visibility of a company's information on the web are rampant and no doubt do offer some tips for doing so. But the reality is that not all search engines search the same sources nor do they do so in the same way. The frequency and method of indexing varies considerably from one search engine to another and from one site to another ranging from almost instantaneous indexing for breaking

news stories and weather, to much slower and sometimes only partial or select indexing of large repositories even when they are made known to the bots (more about this aspect is addressed later in the paper). The same search will return different results from one search engine to another, not only due to content but also due to the ranking of the items displayed to the user. As ranking algorithms are generally proprietary, it is unknown how much weight is given to timeliness of the information, site popularity, data reliability, sponsors and/or advertiser relationships, and other factors. But despite these variables, users are generally highly satisfied with the results from Google especially, and rarely question if there is any information missing as they already get more information than they can review.

More detailed strengths observed in the study were that Google and GS handle plurals better (they find both cell and cells, for example, where *ETDEWEB* currently does not unless the user specifically asks for it) and they also apparently have some level of semantic-based search aids helping the results. It is surmised that Google and GS also return results based on the full text of documents, when available, whereas the *ETDEWEB* searches done for this study mimicked the Easy Search, which does not include the full text (note: *ETDEWEB* does have full text searching available, however, and would have meant even more records were found).

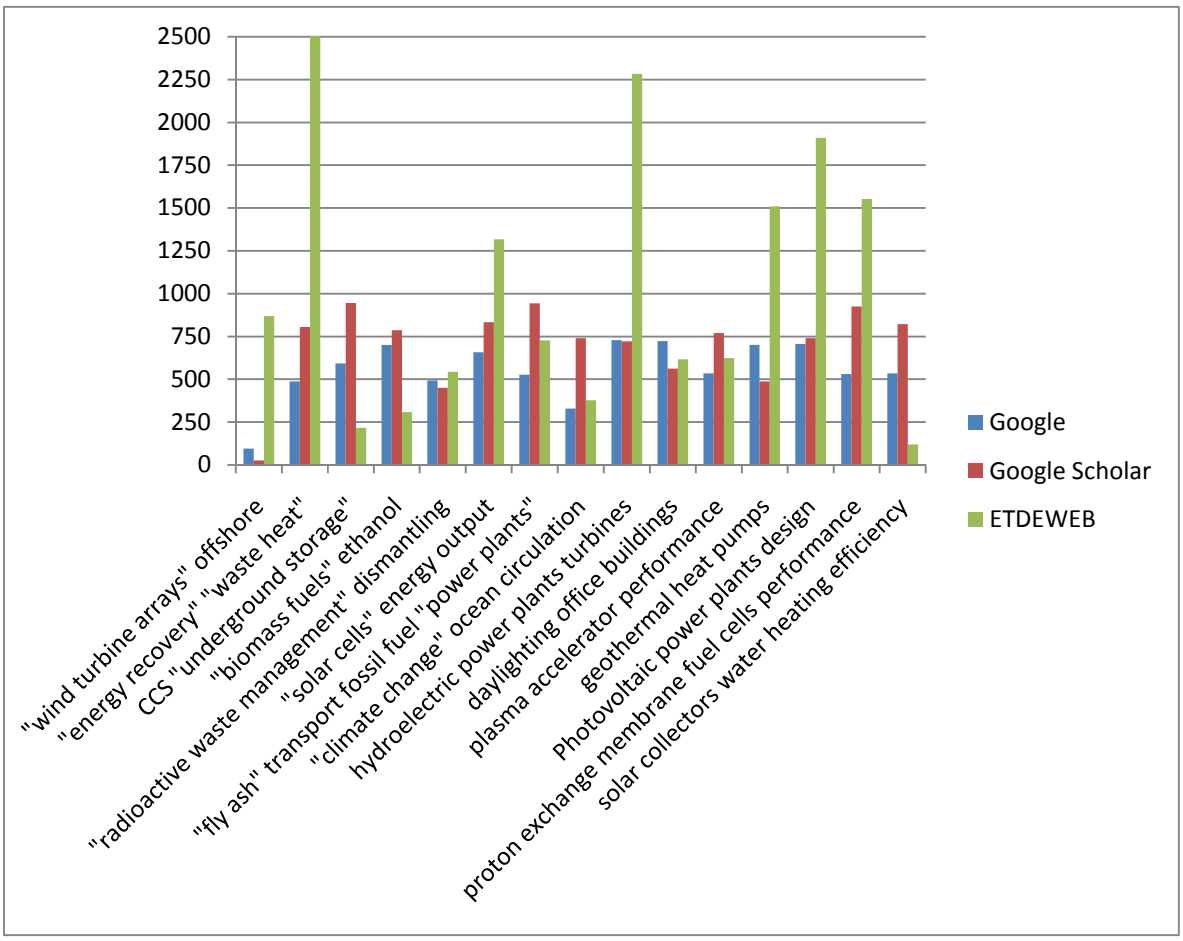
Study details and results

The analysis compared the results of the 15 energy-related queries performed on *ETDEWEB*, Google and GS using the identical words. A variety of subjects was chosen, although mostly in renewable energy areas due to high current interest by many of ETDE's users. The specific terms used were aimed at being specific enough to keep the number of records returned to a manageable level rather than high volumes. The results from each query were captured as 45 data sets with titles and source information that was then used for the comparisons. Over 40,000 records were part of the study.

Impacting the comparisons was an interesting phenomenon observed in the WWS study and found to be true in the *ETDEWEB* study as well. Although the initial result quantities stated by Google and GS (e.g., "results 1-10 of about 658,000") are very impressive, in reality, the maximum number of records a user can actually page through and view never exceeds 1,000. Further, as the user scrolls through page after page, Google and GS refine those quantities (apparently eliminating duplicates or highly similar results), and the final count will almost always be lower than the quantities initially stated. Since many users typically do not even look beyond the first several pages of results, this pattern it is not so evident to the casual user, nor maybe so important. But it did mean that the actual number of records available for comparison in the study was always limited to less than 1,000 for Google and GS results. Once the records were in the datasets, some additional duplication within each result set was observed and these duplicates were eliminated before the final comparisons were made. Table 1 shows the search terms used for the queries and the number of records actually retrieved by each of the search systems with the duplicates removed.

The title of the article or document was the common denominator across the systems and the principal field used for the comparisons. The analysis looked at the source field as well for other factors, but not for direct matching. Greater scrutiny was given to subsets of the *ETDEWEB* records (such as the peer-reviewed journal items) that were expected to be present in the search results from Google and/or GS to ensure the comparisons were accurate.

Table 1. Search terms and result counts



As reflected in data driving table 1, *ETDEWEB* averaged 1,403 records for the searches while Google had 556 and GS, 704, respectively. The higher average results for *ETDEWEB* obviously reflects well on *ETDEWEB*, but it is a bit misleading as an average, given the limits on Google/GS results. *ETDEWEB* had no upper limit on its results returned although a limit of a few thousand was targeted. The initial exact title matching between the sets from each of the queries produced only 57 matches per search, on average, between *ETDEWEB* records and the Google/GS result sets. GS matches were more prevalent than Google matches by a ratio of about 17:1. This ratio was not unexpected, as GS stresses more scientific/technical content. Major sources observed in the actual data for the GS sets included international journal publishers, technical society publications, conference papers, academic sources, and also records from two of OSTI's own databases, Information Bridge and Energy Citations. The Google results were much more varied with contents ranging from blogs, wikis, news articles, press releases, business, government and consumer-oriented sites to patents and books from publishers and product descriptions. One of the results the study expected to see was that at least some of the matches in the Google results sets would be titles whose origin was *ETDEWEB*. Although ETDE has made a significant percentage of its database content available to the Google bot and other crawlers since early 2009, it was disappointingly rare to see an *ETDEWEB* record in the Google sets. Follow-up testing showed that ETDE is still not ranking very high in the results list in the majority of cases when searching Google, or else the records are eliminated as part of the search engine's selection process. A

further review of metrics related to crawler activity in 2010 estimated that Google is indexing less than 10% of the *ETDEWEB* records made available to it. Research is underway to determine any steps that can be taken on ETDE's part to improve the percentage indexed. If ETDE's information is really not all 'out there' despite making it available, how many other collections are only partially indexed?

The next step of the study sampled unmatched *ETDEWEB* results from the international journal publisher Elsevier that were known to be indexed by GS. This in-depth review led to further refinements in the title matching. An additional 129 matches (average) were made after these refinements. Table 2 provides the percentage of *ETDEWEB* results that were unique for the same search in each system after these refinements.

Table 2. Percentage of *ETDEWEB* retrieved documents not retrieved by Google or GS

Search Terms	**Percent Unique to *ETDEWEB* Results Set**
"wind turbine arrays" offshore	90.0%
"energy recovery" "waste heat"	87.3%
CCS "underground storage"	87.1%
"biomass fuels" ethanol	87.8%
"radioactive waste management" dismantling	85.4%
"solar cells" energy output	87.3%
"fly ash" transport fossil fuel "power plants"	77.6%
"climate change" ocean circulation	85.9%
hydroelectric power plants turbines	87.8%
daylighting office buildings	82.3%
plasma accelerator performance	84.5%
geothermal heat pumps	86.4%
photovoltaic power plants design	88.1%
proton exchange membrane fuel cells performance	86.0%
solar collectors water heating efficiency	86.1%
Averages	**86.7%**

Looking next at the remaining unique *ETDEWEB* records, other ways to find the records in Google/GS were attempted to see if the records were indeed 'out there'. In many cases they were retrievable if entering the exact words from the title of the article. But again, these same records were NOT being returned in the Google and GS result sets in the topical queries used in the study. The analysis showed that just because a document is available on the web does not mean that it will be returned in a Google or GS search. Much filtering and ranking occurs, and what the user is shown is suspected to be at least partially based on popularity of the source, as many of the records found weigh heavily toward certain sources. And if records are similar, the 'prestige' source was typically the one displayed. The study concluded that if a user knows a very specific title, all three systems are helpful in finding the user a source for the document. But if the user is looking to discover relevant documents on a specific topic, as was this study's focus, each of the three systems will bring back a considerable volume of data, but different. For the topics in this study, *ETDEWEB* was shown to bring the user unique results not shown by Google or GS 86.7% of the time.

In conclusion, Google certainly gives users a valuable tool to find an abundance of good information, and GS helps the user focus on scientific disciplines. But if a user's interest is scientific and energy-related, *ETDEWEB* continues to hold a strong 'niche' position in the energy RTD information field and adds considerable value in knowledge discovery.