INFORMATION SYSTEM RESEARCH REPORT

David Provost
1 Dec 2020
LIS 640 - Information Organization and Access

PubMed

PubMed is a database of over 30 million citations and abstracts of informational resources in the biomedical sciences, operated by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). PubMed is the public-facing endpoint of an indexing and cataloging that involves a number of different efforts across the NLM. The smallest component of PubMed is Bookshelf -- a full-text index of a limited number of biomedical books, chapters, reports and other documents -- and PubMed also includes the 6 million articles permanently preserved and available to read for free in PubMed Central (PMC). By far the largest component of PubMed is MEDLINE, the NCBI's database of over 26 million citations, pulled from 5,200 journals in over 40 languages and indexed by trained indexers using the Medical Subject Heading (MeSH) thesaurus developed at NLM since the 1950s.

All of these resources are available online for free to the general public, but two-thirds of the users of PubMed are medical professionals and researchers (Lacroix & Mehnert, 2002). Those numbers are from 2002 and it would be interesting to see if that proportion has changed over the years, as the general public has become more familiar with information seeking via the internet. The massive amount of medical literature, and the importance of making that literature accessible and searchable requires a unique and powerful search feature, and the MeSH systems is what enables that easy access to a deep and complicated database.

An update to the user interface of PubMed in the summer of 2020 has simplified and streamlined the basic search as it is presented to the user (Fiorini et al., 2017). A unified

search bar invites a natural language search, but the application of PubMed's Automatic Term Mapping (ATM) controlled vocabulary and authority control behind-the-scenes makes even the most basic search both more powerful and more carefully targeted. ATM takes any search terms that are not specifically tagged by the searcher and applies them to a number of translation tables, a Subject translation table (including MeSH), a Journals translation table, the Author index, and an Investigator index (National Library of Medicine, n.d.). In this way, even searches that use layperson's terminology will get matched to the expertly indexed controlled vocabulary of MeSH. For example, a search for **cancer** will get mapped to the MeSH term **neoplasms**. The other strength of the MeSH database is the way all MeSH terms are organized hierarchically, and (unless otherwise specified by the user) any search that matches a MeSH term will automatically "explode" the search into any terms below the identified MeSH term in the hierarchy. To refer to the previous example, the user searching for **cancer** will not only have the **neoplasms** MeSH term added to their search, but also will have dozens of more specific MeSH terms applied, such as **cysts**, **leukemia**, **bone neoplasms**, etc (Chapman, 2009).

Partially because of this powerful tool, it is important that PubMed offer users the ability to manage their results, and there are indeed several ways to do just that. The most obvious are the filters that populate the left hand sidebar in every page of search results. The "PubMed 2.0" update offers a streamlined and simplified filter interface. Particularly prominent is the date range filter. Given the types of materials indexed in PubMed and the users that will be searching for them, it is likely that limiting results by date will be useful. Some searchers will only be interested in cutting-edge research, while other users may be interested in finding historical, foundational citations in their field, and will only want older results. Other filters include the type of article (randomized clinical trial, systematic review, etc), a language filter, and filters relating to the age, sex, or species (human or other animals) that are included in the article.

There are two additional powerful tools that PubMed offers that take advantage of the metadata attached to each article to provide a connected network of scholarship. The

most obvious is the “Cited By” tool. By using unique article identifiers, PubMed identifies any articles that have cited the article you are looking at, and provides links to them. This offers an easy way to trace an older article to its influence on more current research. Additionally, PubMed uses a word similarity algorithm to find a selection of “Similar Articles” that might highlight related but unlinked articles to provide a serendipitous opportunity to expand your search (National Library of Medicine, n.d.).

These two features, combined with the powerful controlled vocabulary and authority of MeSH suggest a possible direction for a theoretical “PubMed 3.0”. An even more interconnected experience, where the database uses citation analysis and a more refined similarity algorithm to create a sort of “research cloud” that shows how certain articles cluster and create nodes around particularly influential articles, and how different fields of study may or may not intersect. Trujillo and Long (2018) demonstrate how this clustering can work strictly using citation data. By combining that with similarity and MeSH indexing, additional structures may be revealed within the literature. It will be a challenge to present that sort of analysis to the average researcher in a way that they will be able to utilize, but it is potentially worth investigating, especially in a field like medicine, where more effective and efficient research can result in dramatic advances.

References

Chapman, D. (2009). Advanced Search Features of PubMed. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *18*(1), 58–59. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2651214/

Fiorini, N., Lipman, D. J., & Lu, Z. (2017). Towards PubMed 2.0. *ELife*, *6*. https://doi.org/10.7554/eLife.28801

Lacroix, E.-M., & Mehnert, R. (2002). The US National Library of Medicine in the 21st century: Expanding collections, nontraditional formats, new audiences. *Health Information and*

Libraries Journal, *19*(3), 126–132. https://doi.org/10.1046/j.1471-1842.2002.00382.x

National Library of Medicine. (n.d.). *PubMed User Guide*. Help - PubMed. Retrieved November 29, 2020, from https://pubmed.ncbi.nlm.nih.gov/help/

Trujillo, C. M., & Long, T. M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, *4*(1). https://doi.org/10.1126/sciadv.1701130