21st Century Tools for Researchers and Practitioners

Using Automated Tools for Knowledge Curation

Internet search engines have empowered citizens in their quest for seeking insights on a multitude of issues. Knowledge curation and evidence review requires systematic and rigorous fact-finding, baseline subject matter expertise, and the right tool to work at scale. Finding and summarizing knowledge has a direct impact on the research and dissemination of evidence-based practices and novel approaches, and on improved outcomes of interest.

Literature reviews are the most common methodology for knowledge curation but are limited by lack of human resources and the sheer number of publications available. It is estimated that there are approximately 30,000 scientific journals publishing upwards of two million articles every year (Wagner et al., 2021).

In this context, subject matter experts benefit from the support of automated tools to provide customized, iterative, and replicable processes. Many of the world's most challenging problems need solutions that move beyond a Google search. Abt Associates is expert at using our subject matter expertise in combination with automated tools to curate existing knowledge in a more accessible way.

Automated Tools

Text analytics and Natural Language Processing (NLP) tools are being used to modernize and expedite search processes in literature reviews (Qin et al., 2021). These tools augment traditional literature review processes to allow faster and more-sophisticated categorizing, filtering, and searching of large sets of peer-reviewed literature. The algorithms that support data gathering additionally help to build a knowledge infrastructure customized to the research domain, which produces a replicable system tailored for continuous knowledge discovery and curation. These tools also generalize so that web scraping and discovery of gray literature, such as reports and white papers that are not found in peer-reviewed journals, can also be used to contribute to systemic reviews and can help identify emerging topics and initiatives.

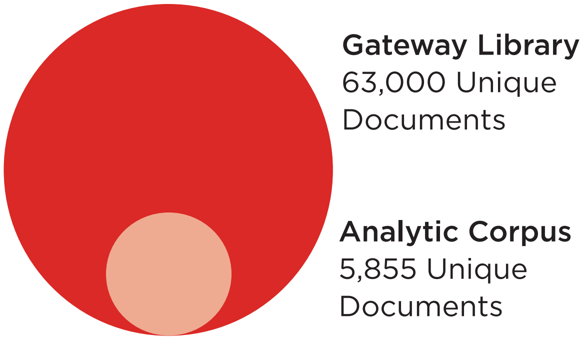
Once data are gathered, researchers can use NLP and Machine Learning (ML) tools to filter documents, model topics, and label documents. Integrated into an ongoing or iterative process, these algorithms can learn to identify possible incorrect labels, flagging them for follow-up by researchers and improving the overall accuracy of the database. These processes, and the work done to standardize and tokenize (removing grammar related features to harmonize a concept such as walk, walks, walking, walked, etc.) text, become the groundwork for more-sophisticated analyses as the number of documents added and reviewed grows over time, e.g., by leveraging complex semantic and grammatical rules derived from massive datasets of millions of documents.

Benefits

- The ability to gather many times more documents than would be humanly possible in a traditional literature review, and to search broadly using webscraping to discover nontraditional knowledge sources or gray literature
- A replicable, adaptable, and repeatable system for managing, monitoring, and maintaining the knowledge management framework to support meta-reviews
- An evolving database of documents that are organized into topics of interest; this can provide inputs for future labeling and categorization of new documents using ML tools, such as user-defined ontologies or NLP-based semantic modeling
- Iterative processes with subject matter experts; these improve algorithmic accuracy over time, which can help identify possible user errors or misclassifications
- Automatic identification of emerging trends or new areas of research
- A process of capturing metadata that can enhance our understanding of researchers and institutions that are driving the research in identified topic areas

Real-World Application

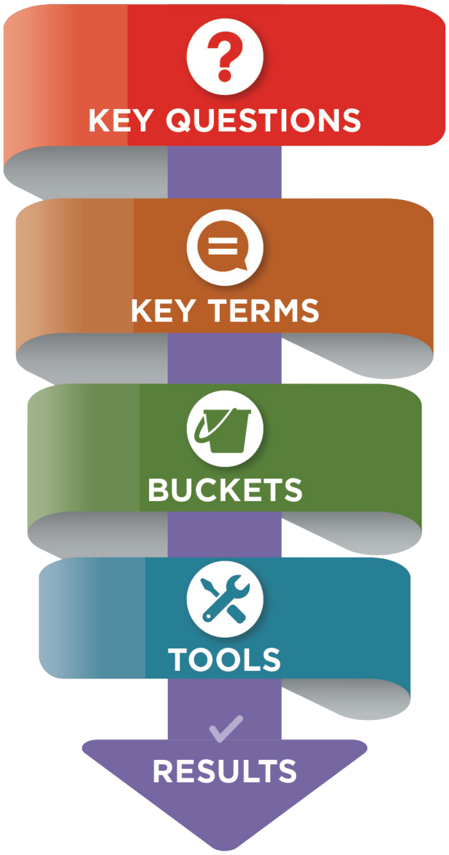
As a result of a research sprint on COVID-19, Abt was interested in applying a similar process to a publicly available website and selected the Child Welfare Information Gateway (CWIG). The child welfare sprint team web-scraped (i.e. use computer algorithms to automatically gather data from a website) and analyzed over 30,000 abstracts and documents on child abuse prevention, adoption, and foster care.



Both sprints resulted in a machine learning powered literature synthesis tools, as well as processes that future teams can use to collect, clean, analyze, and visualize large amounts of text. The promising results of this early sprint led to a pilot in which Abt expanded and applied its processes and search tool to wider data literature synthesis libraries (Cochran and Campbell) and peer-reviewed databases (JSTOR, Academic Search Complete, and Medline) to answer additional key questions

Parallel Processes: Subject Matter Expertise and Technical Applications

As in a traditional literature review, the pilot project initiated the search process through the input of subject matter experts who cocreated the key questions they wanted to answer, and then identified and selected terms aligned with the areas of interest. Where Abt's approach differs from traditional approaches is in deepening and extending the role of subject matter experts to align with our use of modern AI algorithms; we implement a Rapid Synthesis and Translation Process (Wandersman et al., 2008) combined with an Interactive Systems Framework (Thigpen et al., 2012). Using both together we aim to align research to practice and knowledge to action by incorporating phases that influence one another. Once the baseline combined framework is established, we shift our work to the creation of logic statements for search, and filtering documents into buckets of interest. For each of these activities, the subject matter expert and technical teams work in tandem to iteratively check the results, review key terms and logic for grouping ideas and concepts together into clusters, and rerun the tool periodically to capture the newest publications available or add new literature repositories.



The team defined 16 clusters of interest (population pre-natal to age three and pre-natal to age five population other, population maternal, systems funding, systems coordinated service delivery, systems maternal health, systems support services, systems equity, community approaches, community neighborhood, community environmental determinants, community housing, community equity, program support services, programs family centered services), as well as two additional clusters to focus on: family resource centers and mandated reporting. For any given cluster, we apply logic groupings such as "(PN-3 OR PN-5) AND systems equity," to find documents matching those cluster criteria. For the keywords, we use the title/abstract to get papers along with the count of each keyword.

For data processing, we applied tokenization to the keyword lists generated by the subject matter experts and verified that they retained the initial meaning from that team. Once that was confirmed, we used the final version of the tokenized keyword list and the tokenized title/abstract, and then grouped the results by each publication record, so that for each paper we had a count of the number of keyword tokens in the title and abstract. For the analysis step, we focused on a subset of papers that met the following criteria: the paper has (1) at least one keyword in the title or abstract from the target population category, and (2) at least one keyword from the other topic categories (e.g., systems, community, program, or equity).

This group became our analytical set of papers for each literature repository and is referred to as our "broad group corpora," which we further narrowed down. Each phase built incrementally upon the prior results and iterative interpretation by the subject matter team and identified stakeholders, resulting in a smaller subset of literature more closely aligned with the key terms.

Below is a summary of the number of publications in the broad group as well as their percentages out of the initial count that were included in our broad group corpora.

Percentage of Publications Included in Broad Group Corpora

Repository	Number of Publications	Publication Year Range	Broad Group Corpora
CWIG	63,416	2000-2021	7,077 (11%)
Cochrane	8,650	1997-2021	1,424 (16%)
Campbell	569	2005-2021	30 (5%)
JSTOR	15,694	2000-2021	154 (1%)
Medline	98,895	2000-2022	5,270 (5%)
Academic Search Complete	158,867	2000-2022	4,086 (3%)
Total	**338,306**	**1997-2022**	**18,041 (5%)**

Ways to Explore and Illustrate the Data

Part of the value of conducting text analysis is that once data are structured into a format that are easy for computer algorithms to digest and provide insights, we can look at summary statistics such as search term frequency (Figure 1), review machine-generated keyword lists or topics to see how they compare to our topics, or analyze co-occurrence (Figure 2) of keywords to understand how terminology is used in the body of literature. This can help us revise or refine our search techniques as well as guide our thinking in subsequent iterations.

Figure 1: Raw Frequency of Search Terms

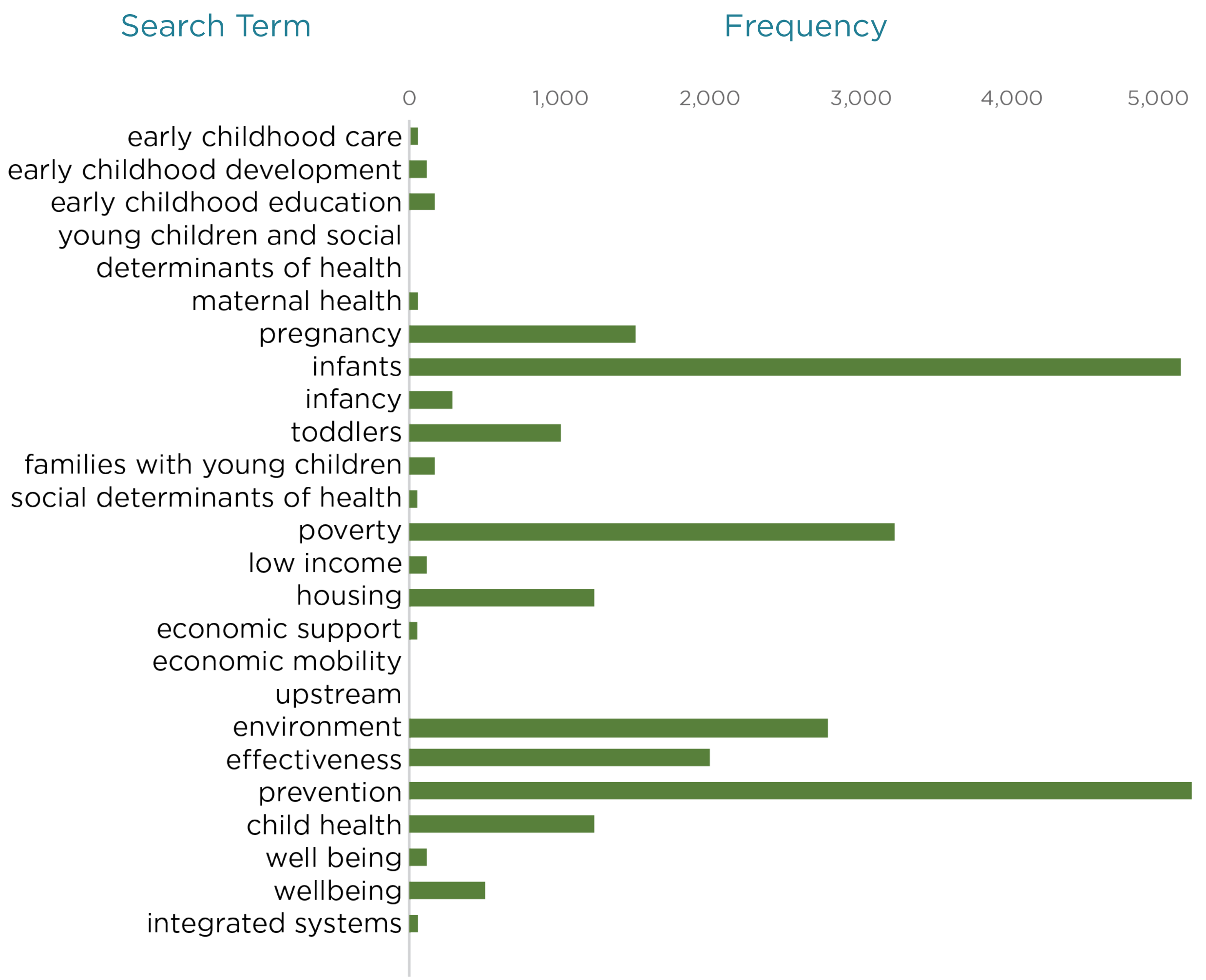
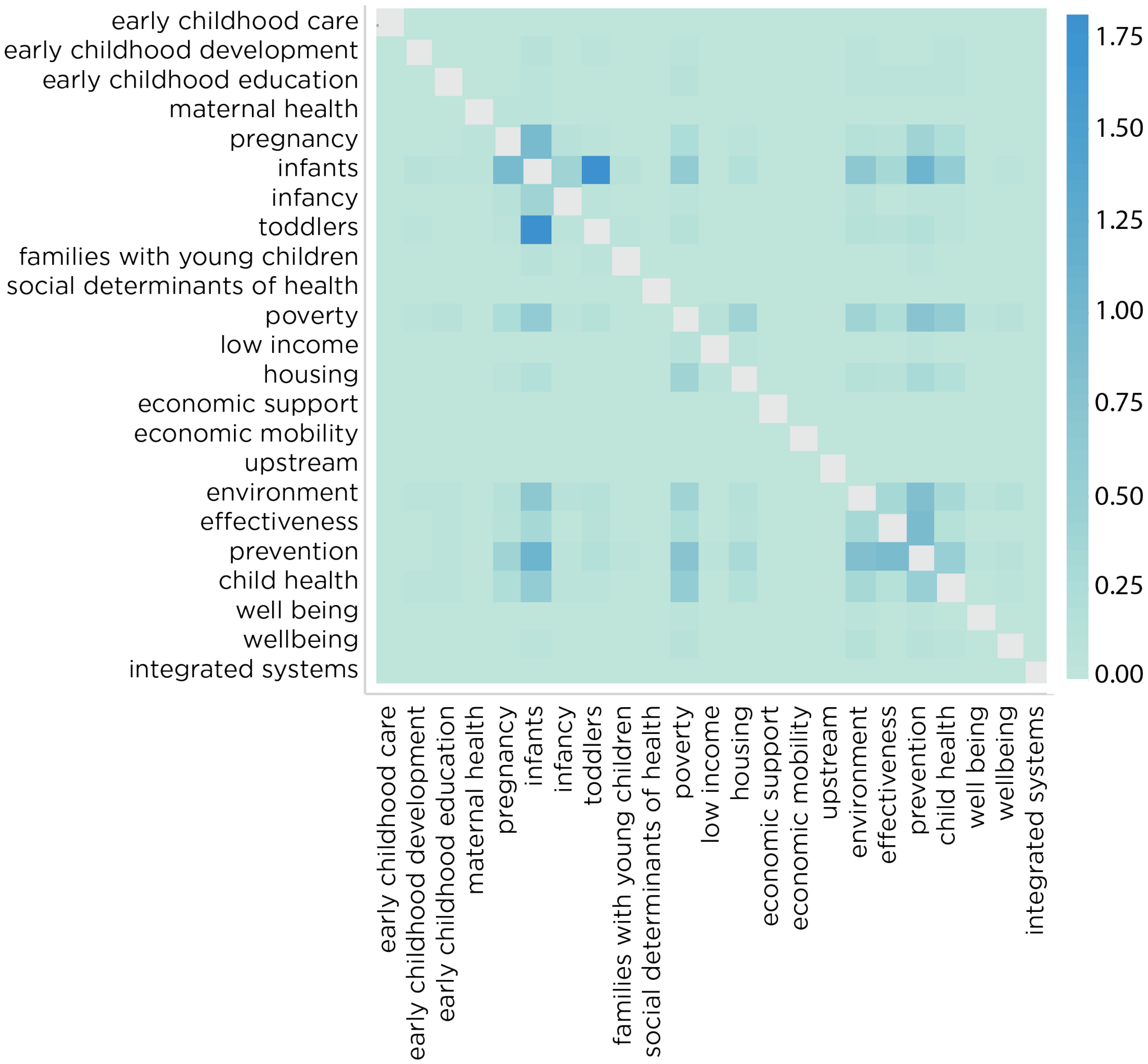
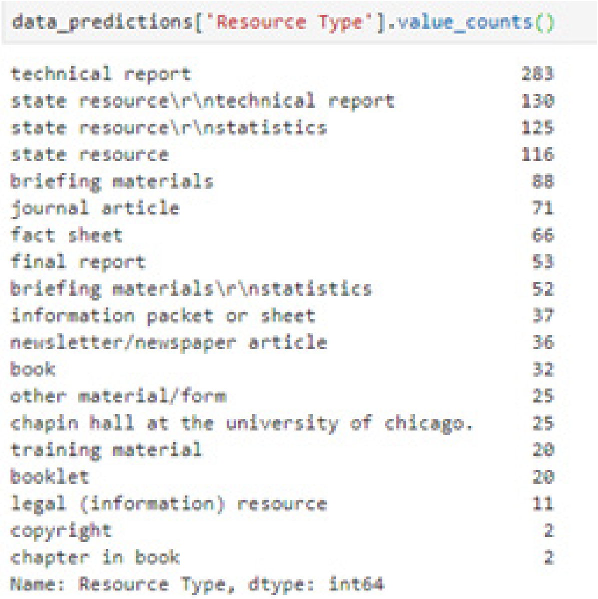


Figure 2: CWIG Keyword Co-Occurence



Automated Labeling and Categorization of New Data

As we document sources and types of data, we can expand into using NLP or predictive tools to label new sets of information. For example, the figure below shows the counts of document types in the dataset that were extracted from the CWIG using the term "foster care." Human tagged keywords from websites, libraries, or literature repositories can be utilized in supervised learning approaches to "teach" AI tools to automatically tag documents without labels into these groups. These document types have similar properties but may not always be labeled appropriately when found through web scraping or document libraries.



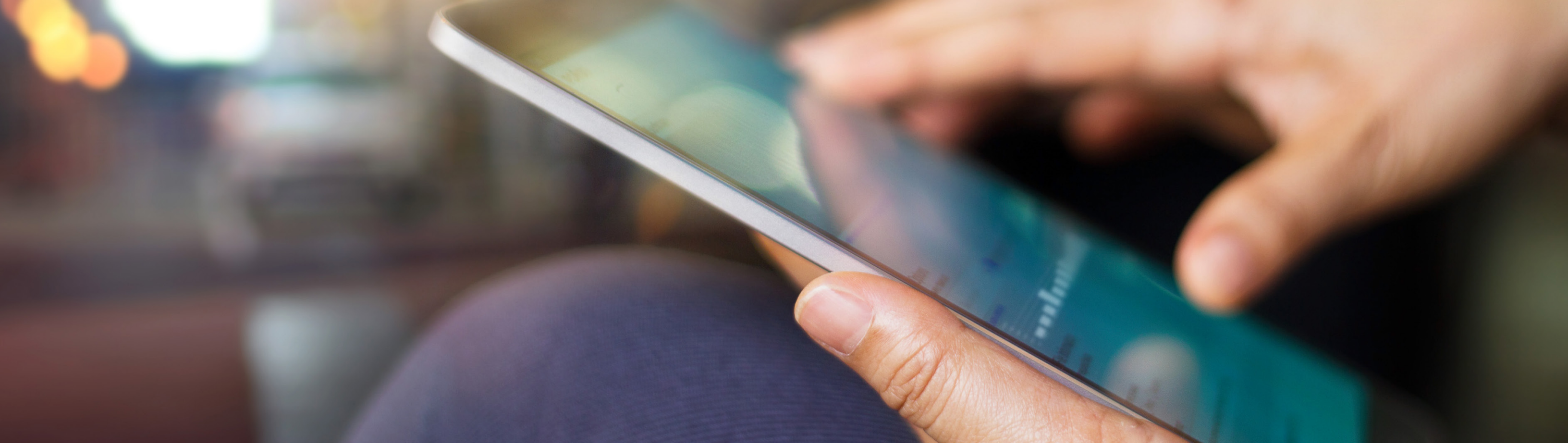
```
data_predictions['Resource Type'].value_counts()

technical report                               283
state resource\r\ntechnical report             130
state resource\r\nstatistics                   125
state resource                                 116
briefing materials                              88
journal article                                 71
fact sheet                                      66
final report                                    53
briefing materials\r\nstatistics                52
information packet or sheet                     37
newsletter/newspaper article                    36
book                                            32
other material/form                             25
chapin hall at the university of chicago.       25
training material                               20
booklet                                         20
legal (information) resource                    11
copyright                                        2
chapter in book                                  2
Name: Resource Type, dtype: int64
```

We can use the structure and textual characteristics of these documents to help us label new documents with suggested types as we add them to our process, by training our tool with the data provided by manually classified databases. Additional labels and categorization can be added to the database at any time. Given a small number of human-tagged documents for training, the ML algorithm can automatically assign tags based on methodology, geography, demographics, or any other human-defined category.

Summary

Automation-enhanced literature searches can begin to solve the problem of examining and exploring vast amounts of information within a topical domain and filtering it quickly to narrow and identify the topic areas of interest. Through iteration with subject matter experts, tools like these can help accelerate and expand the search for published and unpublished works and identify key themes and emerging trends in the literature. These tools accelerate our ability to get highly targeted and evidence-based knowledge into the hands of the stakeholders to efficiently improve outcomes of interest.



References

Wagner, Gerit & Lukyanenko, Roman & Pare, Guy. (2021). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*. 1-18. 10.1177/02683962211048201.

Qin, Xuan & Liu, Jiali & Wang, Yuning & Liu, Yanmei & Deng, Ke & Ma, Yu & Zou, Kang & Li, Ling & Sun, Xin. (2021). Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of Clinical Epidemiology*. 133. 10.1016/j.jclinepi.2021.01.010.

Thigpen, S., Puddy, R. W., Singer, H. H., & Hall, D. M. (2012). Moving knowledge into action: developing the rapid synthesis and translation process within the interactive systems framework. *American journal of community psychology*, 50(3-4), 285–294. https://doi.org/10.1007/s10464-012-9537-3

Wandersman, A., Duffy, J., Flasphor, P., Noonan, R., Lubell, K.,Stillman, L., et al. (2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. *American Journal of Community Psychology*, 41, 3–4

For More Information

Christine Tappan, MSW, CAGS
Principal Associate, Health, Social & Economic Policy
Co-Director, Abt Global Center for Technical Assistance and Implementation
Telephone: 410.693.4342
Email: Christine_Tappan@abtassoc.com