Copyright Clearance Center | RightsDirect

Whitepaper

Beyond Standard Search

Getting the Targeted Data Your Organization Needs

Sometimes the data an organization needs to make critical business decisions is simply not available, or at least not ready and waiting. Sometimes a dataset needs to be created before a company can begin to get insights that will help inform its strategy. And sometimes teams need to scale their processes around gathering and analyzing data without increasing manual effort — data that may itself be time-sensitive.

The information that companies need to make the right decisions can vary widely, but clinical trials, technology transfer offices, patent, conference, and competitive company information, and global regulatory, corporate, and organization news all offer data from which insights can be derived.

In this white paper, we explore the value that web crawling, creation of curated datasets, and delivery of targeted and relevant intelligence can provide to an entire organization, including data scientists, information managers, competitive intelligence professionals, and business development teams seeking information that will help them make the right decisions for their organization.

Finding Crucial Data in the Gaps

Organizations are increasingly seeking to make data-driven decisions, as doing so can:

Contribute to revenue growth

Improve productivity

Reduce costs

Streamline product development

Strengthen understanding of the competitive landscape

The Limitations of Search Engines

Relying exclusively on common search engines such as Google can make it difficult to find vital data in the gaps.

A gap, for example, might be key documents or pages that aren't indexed by popular search engines. Additionally, search engines commonly organize the results they present by what most users are searching for, so an engine such as Google uses its algorithm ("what do most people who are searching on this term want to see?") to create a common answer.

But what if an organization is seeking an uncommon answer? To find it, a search might need to go well beyond the first few pages of results presented, potentially leading to a significant investment of time. Rather than use search engines that may apply an undesirable context to results, companies are better served controlling how their search results are presented by being able to apply their own context to their search.

To become data-driven effectively, organizations naturally seek more and better data. Responsible organizations must consider where the data that supports their decisions and analyses is sourced. Typical sources include the databases used by many businesses, including those for published literature and grants, as well as regulatory databases.

At CCC, we've discovered from our work with clients that desirable data for making better-informed decisions can exist outside these common sources — this data is found in "the gaps." While licensed or structured datasets can be beneficial, they do not cover everything and sometimes an organization needs to fill the gaps between their data sources to create a single source that is unique and particularly relevant to the organization's needs. When everyone is licensing and subscribing to the same datasets, a competitive advantage can come from using information somewhere in the middle, or in the gaps.

One example of data found in the gaps pertains to monitoring social media; organizations are interested in what others are saying about their products, and they are also interested in what their competitors are saying about themselves in the form of company news and press releases. Beyond social media, "gap data" can also be found in obscure, but critical, areas such as tracking product shipments across borders.

Crawl, Enrich, and Deliver with Deep Search

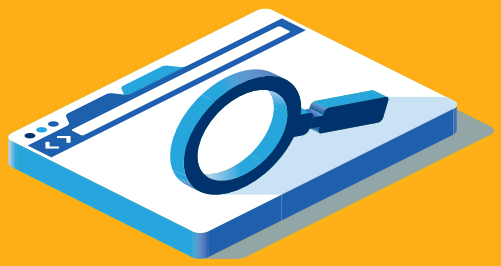
While utilizing data found in the gaps can be hugely beneficial, there are challenges to doing so. This data likely hasn't been tagged, curated, or normalized, leaving it unstructured and distributed and making the process to collect it manually intensive. Further, the information organizations are seeking can often lie alongside data not relevant to the context or need, leading to a lot of noise to filter out. Gathering relevant data in the gaps might bring in human error, then, as organizations assign workers who might not have deep knowledge of the business to sift through data.

Organizations seeking to successfully locate and utilize data found in the gaps have a solution they can turn to: "deep search." Deep search is a solution encompassing a set of tools and processes that, subject to the terms of access and use of the relevant data source, can crawl targeted sources of data, enrich them, link and curate them, and then deliver them to a focused audience. A deep search approach can help organizations get more value from their existing datasets, help them create new datasets from one or more sources, and even assist them in efficiently scaling processes and quickly collecting time-sensitive data with reduced manual effort.

Getting More Value from Your Data

Organizations using licensed or publicly available databases may not find all the data they need in these sources, or they may encounter the opposite problem, finding an overwhelming amount of data full of noise that's distracting and difficult to sift through as they try to locate the information they need.

A common example of the above situation can be seen in researching grants and funders. Organizations interested in tracking grants of interest, and understanding more about the funders behind these offers, are often trying to better position themselves to receive them. There is also simple intelligence to be had from seeing what grants are out there, and who is funding what. Grant tracking can be hampered by potential issues, including gathering information that is irrelevant to the organization's needs, perhaps because a broad database is used covering domain areas of no interest, or too large of a geographical region. In other situations, the organization can fail to obtain the data it needs, as when a database does not include relevant information, such as the link between a principal investigator and their institution, that would help an organization better understand the context around a grant or funding.



By having to either tune out the noise of unnecessary data or find additional data sources to get the information needed, organizations can lose valuable return on time invested. On top of this, the need to use multiple grant databases makes it highly unlikely that the data found will be consistently normalized, increasing the difficulty in getting a complete picture.

By having to either tune out the noise of unnecessary data or find additional data sources to get the information needed, organizations can lose valuable return on time invested.

Deep search helps address the above issues in part by using database information as a starting point, and, when multiple databases are used, cleaning and normalizing this data. Potential noise is automatically tuned out through deep search by an organization's ability to focus on what it wants to see through filtering, keywords, ontologies, etc. By helping connect an organization to relevant information, a more complete, targeted dataset is created that focuses exclusively on the area of interest extracted from the wealth of available data.

In this context, deep search can help an organization maximize its return on time invested and the value of the licensed data sources it may have paid for. This is also true when your licensed databases might not have certain information — deep search makes it possible to add new information to a combined dataset by crawling websites in a focused, targeted way. The information obtained through deep search in this way can then be shared with stakeholders via emailed reports, and alerts can be set up in case information changes.

Solving Problems with New Datasets from Multiple Sources

Of the many applications of deep search benefitting organizations, a standout is the ability for an organization to use the solution to create an entirely new dataset when no single source exists and essential information is spread across multiple unconnected sources.

Deep search can help an organization maximize its return on time invested and the value of the licensed data sources it may have paid for.

Creating a Dataset When No Single Source Exists

To illustrate how deep search can create databases from essentially anything, provided that the information needed can be crawled both legally and technically, we consider an example of an organization needing to track activity around regulations to inform its product strategy for both existing products and those in the research phase. The specific information sought by the organization is not contained in regulatory documents, but in some of the comments made on dockets.

For context, regulation documents on a site such as regulations.gov can each receive thousands or tens of thousands of comments apiece. Without using deep search, an organization seeking the desired information would have to review all the documents on this site, filter for the ones related to its product space, read through thousands of comments, and then provide a synopsis of what was said in the comments relevant to the organization's interests.

Alternatively, this same organization could employ deep search to sift through the thousands of comments, identify the ones relevant to the organization's needs by filtering for and applying key words, and then direct only the contextually relevant comments to specialists for review, with the insights they gain helping to inform product strategy. As seen in this example, when desired information does not exist as a single source or as a database in and of itself, deep search makes it possible to collect and curate this information efficiently so that it can be interrogated for useful insights.

Gathering Information from Unconnected Sources

Data collection from unexpected and unconnected sources can be useful to organizations for a myriad of reasons, including protecting their business interests outside of research and development.

A specific example of this involved a client of ours who sought to track the movement of their product across borders to ensure that the product intended for one country did not (by unsanctioned import or export) end up in another country where pricing may be different. While this sort of activity is not uncommon, it can be hard to track and is usually discovered by chance.

For our client, the data it sought did not readily exist in a single source. Before we helped them with our deep search solution, its process was ad hoc and required a great deal of time and manual effort, including digging through several different data sources such as TradeAtlas and checking customs bills of lading to track the progress of the product across borders.

By using deep search, the client was able to create a single data source compiling all necessary information from different licensed databases. Keywords could then be run across this dataset to identify the company's product and track movements across borders. Additionally, information in this dataset could be de-duplicated and normalized to provide a 'clean' view of only relevant data that could then be shared and acted on. Through a deep search approach, the client can now manage the volume of the above work with only a small team.

Efficiently Scaling Processes and Collecting Urgent Data

Utilizing deep search can help an organization reduce the manual effort needed to secure crucial data found in gaps, such as across multiple unconnected data sources. The solution can also help companies who need to gather data quickly to help inform time-sensitive decisions.

Scaling Up Manual Curation with Less Effort

Even when sources of information are already available to an organization, deep search can be an effective solution for reducing the manual labor needed to find relevant data. In one example, a client of ours tracked company news as part of gathering competitive intelligence. The sources of information for this client were easy to locate and included competitor websites. The existing process the client used to track competitive intelligence involved a single employee checking these various websites for relevant content, including press releases, investor news, company newsletters, etc. The employee then summarized the content manually.

When applying deep search to this organizational need, the client was able to significantly increase the number of websites tracked, as well as index and tag found content to an internal ontology so relevant information could be searched and delivered to different business users. All this tracking was completed three times a day to help ensure the organization had access to the most up to date information available.

The key information that was tracked and rated was then pulled into a weekly report that was automatically emailed to the target audience, saving even more manual effort. Though the project discussed here did not involve a complicated process, it was still one that was onerous for a single employee to take on. By employing a deep search solution, the client was able to scale this process and create many more useful competitive intelligence insights.

The Benefits of Using a Trained Model

Clients like the one discussed here get even more from their deep search solution by using a trained model — a form of AI or deep learning that requires an initial set of data which domain experts can look through and classify with any labels they choose.

In this instance, the client was able to rate the importance of each item tracked using a trained model, with documents in the initial dataset labeled as "important" and "not important." This dataset established a baseline and trained a model through machine learning, and the model was then applied to incoming data sourced via crawling, allowing these newly found documents to be labeled automatically.

A mechanism was set up for newly crawled information, allowing users to evaluate the automatically applied rating and override it as necessary. This feedback loop enables the client's machine-learning model to be updated on the fly and helps improve its accuracy over time.

Deep search even allows for following internal links to deeper information (e.g. company websites referenced in the abstracts) and adding that information into the mix, another example of extracting value from data found within the gaps of intersecting sources.

Seizing on a Small Window of Opportunity

There are also instances where deep search can benefit an organization when there are limitations imposed on data outside the normal expectations, such as when the data is transient and time-limited. We have seen problems related to this type of limitation arise frequently, and often the solution requires access to new information the instant it's published, with any delay resulting in missed opportunities for an organization. To illustrate this, we can take the instance of a company seeking to extract value from pre-conference activity. This type of information is updated regularly, even as frequently as hourly in some cases.

One benefit of this information includes the ability to sift through conference exhibitors, find those of interest, and research their websites to help make an organization's attendance extremely targeted. Another benefit is obtaining a sense of key conference themes by trawling through all the abstracts to isolate those of interest. The abstracts might even tell a story about where the market is heading, one more benefit to possessing this information.

However, to utilize the most accurate, updated version of pre-conference activity, an organization would need to dedicate employee time to frequently visiting the conference website, trawling the abstracts, trying to identify people, issues, and talks of relevance, and then packaging this information and sharing it internally, a time-consuming process prone to human error.

By using deep search, this process can be automated and made more accurate: crawling across the conference site as often as needed (when crawling is permitted), gathering and automatically curating discovered content, even running an ontology across any findings to focus on topics of interest or key word extraction, all possible as part of a deep search solution. Deep search even allows for following internal links to deeper information (e.g. company websites referenced in the abstracts) and adding that information into the mix, another example of extracting value from data found within the gaps of intersecting sources. In this case, the newly created dataset unites conference information with company and even product information in one place.

The potential of a data source like this one includes:

- Allowing customers to find key opinion leaders to engage with at the conference
- Enabling a business leader to whittle down a long list of exhibiting companies to only those of probable value
- Furthering a company's strategy by identifying target companies for acquisition
- Helping an organization to infer market direction based on a summary view across all abstracts

After the conference, this information source can continue to provide value by allowing an organization to keep an archive of everything crawled and enriched, helping it to continually grow its understanding of the market and adding to its competitive intelligence data. While this example is focused on conferences, a deep search solution can potentially provide as much or more value when applied to any data that's time bound.

Carl Robinson

Senior Corporate Solutions Director, CCC

Carl joined CCC in 2013 in a role supporting publishers with their content management needs and progressed to develop a team of senior consultants focused on publishing and workflow solutions. In his current role for the company, he focuses on research and development workflows. Prior to joining CCC, Carl spent 20+ years working in senior positions at blue chip companies, including Oxford University Press, Macmillan Education, and Pearson.

Stephen Howe

Principal Product Manager for Data and Analytics, CCC

Raised in a family of teachers, a former teacher himself, and a lifelong student, Stephen's goal is to help people learn. Stephen has spent his career working at the intersection of publishing, education, and technology, holding positions in sales, sales management, production, project management, digital publishing, digital editorial, and product management. Trained in the liberal arts tradition, Stephen holds a BA and MA in philosophy, an MBA in management, and a Masters in Analytics.

Learn about CCC Deep Search Solutions

Our deep search solutions offer automated capabilities that gather, enrich, and deliver targeted intelligence from online sources to meet the needs of your organization or department. Gain a sharp competitive edge with highly relevant intelligence from information sources you specify. Learn more at: copyright.com/solutions-deep-search/

About CCC
A pioneer in voluntary collective licensing, CCC advances copyright, accelerates knowledge, and powers innovation. With expertise in copyright, data quality, data analytics, and FAIR data implementations, CCC and its subsidiary RightsDirect collaborate with stakeholders on innovative solutions to harness the power of data and AI.

Learn more about our licensing, content, and data solutions:

U.S. organizations:
copyright.com
solutions@copyright.com

Outside U.S. organizations:
rightsdirect.com
solutions@rightsdirect.com