Copyright Clearance Center | RightsDirect | Imbue PARTNERS

Whitepaper

Accessing and Analyzing Relevant Content in Today's Information Chaos

R&D Challenges and Opportunities

Decreasing time to market can be a competitive advantage for companies across any number of industries — from pharmaceuticals and medical technology to food, chemicals, energy, and more. Knowledge management is essential for corralling and analyzing the massive amounts of information needed to facilitate the work of R&D scientists, researchers, and developers. Regulatory, Safety, Scientific and Medical Affairs, and Competitive Intelligence are additional functions that also can benefit from easier, timely access to critical data. One way of accelerating the early stages of a product's lifecycle, particularly in R&D organizations, is to ensure information users can quickly find and access relevant and accurate data and other content that facilitates discovery and development.

Four core requirements are necessary to realize the full potential of streamlined search, usage, and insight generation.

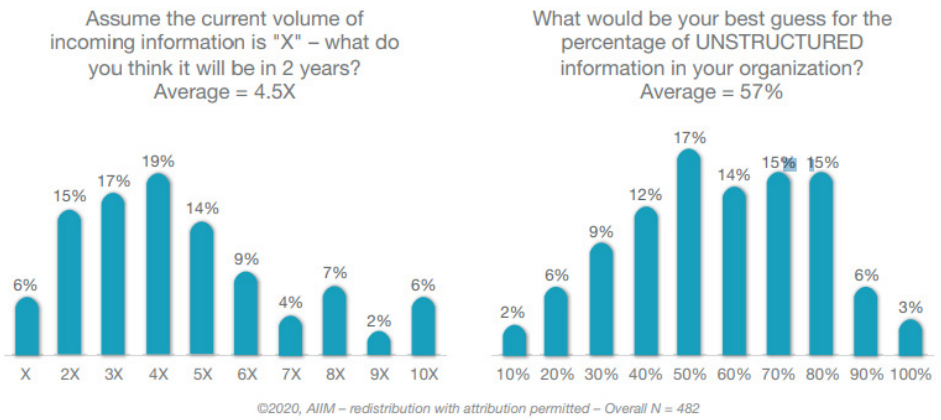
1. Automated integration of external and internal information
2. Elimination of data silos
3. Connection between concepts
4. Organizational development such as hiring new talent, embedding strategic capabilities, aligning leadership, and implementing new processes and tools

Imbue Partners, in collaboration with Copyright Clearance Center, has set out to explore four key functional areas that are most helpful in achieving the objectives above:

- Personalization
- Content aggregation
- Data analytics and visualization
- Semantic enrichment

What makes this difficult? Scientific content discovery and information management can be complicated by the increasing **volume, velocity, and variety** of newly generated information. The pace of freshly available data and the increasing scope of relevant information make it difficult for users to sift through volumes of material in a streamlined, efficient way – especially when driven by time constraints. COVID-19 also compounded the effects of information chaos through the immediate and then enduring effects of knowledge workers conducting their work while offsite, rather than in an office environment.

According to a 2021 AIIM report, the volume of incoming information is estimated to grow 4.5 times over the next two years. Yet, currently, more than half (57%) of the information at the organizations AIIM surveyed is still considered unstructured, which hinders users' from easily leveraging these companies' knowledge and intellectual assets. In the same AIIM survey, respondents were asked to how they felt about organizational progress in the battle against "information chaos." The average grade was a subpar C- with 25% of the respondents representing a line of business and >50% representing content management or similar (AIIM, 2021).



Personalized content

As consumers, we experience personalization daily via targeted online advertising while browsing or in our social media accounts. We see it in the shows and movies that Netflix and Amazon suggest to us and in the music that Spotify or Pandora recommends while we scroll through our playlists. This technology is based on our past purchase and usage behaviors. What we have liked, disliked, who we follow, and what we have previously searched for and ultimately purchased are used by machine learning to predict — and suggest — what might be of interest to us in the future. Let's consider how personalization techniques impact scientific content discovery and information management. According to the latest data, approximately 8.5 billion searches are conducted each day on Google. (1 Second - Internet Live Stats, n.d.). Google — the tool, the term, and the technology — is omnipresent in our collective global culture. This has caused a fundamental change in what individuals expect from search results. Our private life search habits and expectations have unsurprisingly spilled over into our business life.

Where it comes into play

With this shift in expectations, it's important to understand the different types of personalization used by search engines so we can recognize the benefits of applying these tools to data searches in the business environment as well.

When users are searching through and finding content, personalization allows them to find relevant content faster by moving artificial intelligence (AI)-informed recommendations to the top. With **explicit personalization**, search results are driven by a user's chosen preferences, such as setting specific data source selection and/or setting source "favorites." Based on these choices, the user expects more relevant search results to appear. **Implicit personalization** delivers personalized content recommendations based upon a user's past actions and behaviors.

Research indicates 75% of people will never scroll past the first page on a Google search, drastically limiting the range of potential information (Dean, 2020). In light of this data, it is more important than ever that the information most highly relevant to the individual researcher appears at the top of search results.

Challenges and opportunities

At R&D intensive companies, the questions that researchers and other employees attempt to answer are far more complex than a simple Google query can answer. For example, what is most relevant to a researcher working on a promising early-stage drug candidate for Fibrodysplasia ossificans progressive — otherwise known as FOP or Stoneman's Disease, which is expected to affect only 4,000 individuals worldwide — is quite different from what that the same researcher would find valuable in the mature diabetes market.

Why? The Rare Disease field is known for its small patient populations, premature disease understanding, and overwhelming lack of education. Comprehensive and relevant information may be very difficult to find and would require **content discovery solutions** that scour scientific literature, patent information, real-world evidence, and patients' lived experiences from as many sources as possible, including scientific societies and congresses, scholarly publications, clinical trials, social media platforms, etc. Casting as broad a net as possible would help generate novel insights and new discoveries and drive results in this market.

> R&D professionals need software solutions that are based on the right kind of machine learning — in particular, implicit and explicit personalization tools that better "understand" a user's goals.

In contrast, let's look at the diabetes market. Diabetes was accurately described for the first time in the 2nd century A.D.; by January 1922, the first insulin injection was given to a 14-year-old boy dying of the disease (Karamanou, 2016). For the last 100 years, diabetes has been studied by countless principal investigators, labs, and drug development companies. The sheer volume of clinical and observational data and content is vast, and as a result, this researcher's challenge becomes one of technical relevancy and prioritization. R&D users in the diabetes field need solutions and methods to narrow down and contextualize information, to manage the deluge of data, and to help them recognize newly established patterns and trends.

Relevancy in scientific, medical, and technology search. An article's median half-life (more than one-half its total downloads) across all publishers was between 2 and 4 years (Bohannon, 2013). This can bias traditional search engines to favor older publications because citations, impact factor, etc. can take years to develop — missing the mark in identifying potentially novel discoveries and innovation (AIIM, 2021).

This leads us to recognize why R&D professionals need software solutions that are based on the right kind of machine learning — in particular, implicit and explicit personalization tools that better "understand" a user's goals and result in more serviceable content discovery, regardless of the lifespan of a significant scientific paper. Using the right machine learning tool will combine implicit and explicit personalization with the right content to find relevant results.

One user of CCC's RightFind Navigate at a global pharmaceutical company recognizes the value of creating a unified search experience from disparate, siloed content from trusted internal and external sources, saying, "RightFind offers a tremendous benefit as a place to go when you don't know where to start."

Breakthroughs in cancer treatments, rare diseases, and rocket science are possible, not because individual experts know everything on the subject but because people can draw knowledge that does not reside in their own heads (Ward, 2013) — which makes finding the most relevant information at the right time so critical to drive innovation.

Content aggregation

Content host aggregators simply supply information directly from various sources into one feed or display. Familiar examples would be Google News or Apple News, which pull articles from online media sources such as Reuters, ESPN, and NPR into a single page of headlines. For researchers and other knowledge workers using aggregators, gathering relevant data effectively can be overwhelming, given the breadth of material and sources available. In these scenarios, researchers and their teams must comb through the immense volume and variety of data to curate, understand, and apply the insights.

Content aggregators fall into three distinct categories.

1. **Content Hosts:** Companies whose primary focus is to provide a hosting service for publishers. Quality and validity can be challenging as these hosts are not typically selective of the content displayed.
2. **Gateway Hosts:** Companies that index or categorize disparate content on other content host services. These services are generally subscription based and host a collection of links to publishers' full-text content by accumulating abstracts and indexing information. Libraries typically fall into this category.
3. **Full-Text Aggregators:** "Traditional" aggregators of licensed full-text content that encompass everything content and gateway hosts have, but without the subscriptions or creating full-text databases. They centralize access to full content licensing from the original source, which contributes to the validity of the sources and information being provided.

Content curation, in addition to aggregation, provides another layer of expertise which leads to greater relevancy and trust in the search results. In particular, this capability provides users with just-in-time decision support.

Where it comes into play

Organizations looking to acquire a content aggregation and curation service should develop criteria to determine which service would best help their users, while considering the following strategic questions:

- How can we help drive decision making in a timely manner with well-organized empirical data?
- How can aggregated content efficiently become curated content?

> We're allowing new information stores to proliferate and not providing guidance on how to choose the right tool for a particular job.
>
> AIIM 2021 survey respondent

Users are looking to optimize their workflow without jeopardizing the efficacy of search or losing relevant data in the process. Knowledge workers must be able to pull empirical information that directly relates to the research

- to **amplify efficiency**, allowing them to "fail fast" and start down a new path,
- to **spark inspiration and new ideas**, making connections from previously conducted documented work, and
- to **navigate their analysis** seamlessly based on relevance.

Challenges and opportunities

While Google misses both an organization's internal data and other licensed sources — and includes a lot of noise in its results, a more effective search tool employs both aggregation and personalization which work together to strike the right balance of relevant results. Working in tandem, aggregation and personalization provide a researcher with comprehensive information that isn't overwhelming so they can more easily find content that's relevant to their search.

Semantic enrichment

Semantic enrichment is the ingredient behind getting relevant search results even if they don't use the same terminology as the query. For example, a query for "rare disease drug approval" would include results for the Orphan Drug Act from the FDA. Google recognizes that "drug approval" relates to "government regulations." It also knows "orphan drug" and "rare disease" are associated, though different terms are used.

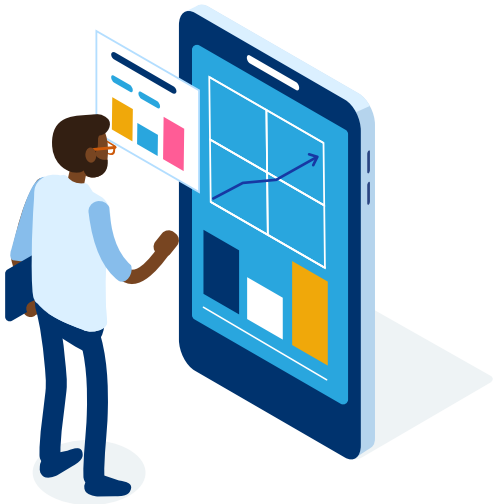
Compare this to another scenario. You've been asked to pull out an important piece of information that was emailed. You scour all your emails but cannot recall the exact verbiage or phrase in the subject line. The email's text-based search function is unlikely to return the correct result unless you use the precise word, which inevitably leads to multiple search attempts and time lost hunting through emails.

The vast differences in algorithms between our two examples — Google and a simple email search — show the power and utility of semantic enrichment in our daily lives. We've grown to rely on search tools to automatically include appropriate synonyms.

Where it comes into play

To eliminate the noise and provide relevant search results, information solutions must go beyond simple keyword matching and to use search engines and algorithms that link concepts, topics, and associations to form a deeper understanding of a user's intent.

For instance, a researcher in pharmacovigilance may need to identify and list all potential Injection Site Reactions (ISRs) before an upcoming clinical trial. Searching published materials might identify traditional symptoms such as sore arm, redness, and inflammation. However, without integrating the company's Adverse Event or Safety database, the search results could miss other unknown reactions such as itching, eczema, and hives.



To tap into external and internal data sources, it becomes necessary to use biomedical vocabularies and ontologies (e.g., NIH's MeSH [MeSH Browser, n.d.]) which are semantically enriched and indexed. The result would be that a search for "Injection Site Reactions" could produce results from known ISRs that had been published previously and catalogued and could also draw from adverse events gleaned through internal sources. A comprehensive solution would account for a company's particular ontology as well as the various vocabularies specific to different organizations within the company.

Challenges and opportunities

While Google continues to evolve its search algorithms, biomedical research has its own set of challenges as noted in the article, "Dug: A Semantic Search Engine Leveraging Peer-Reviewed Knowledge to Span Biomedical Data Repositories" by Waldrop et al: "Despite the practical utility of Google's proprietary knowledge graph for general search, the provenance, depth, and quality of its biomedically relevant connections are not easily verifiable. There remains a need for a search tool capable of leveraging evidence-based biological connections to show researchers datasets useful for hypothesis generation or scientific support." (2022)

This is where functionality beyond linking key terms evolves into topic-linking (or topic co-occurrence). Like Dug, scientific communities and commercial entities are collaborating to improve semantic search. Continuing to build dictionaries and structures to organize, link, and catalog scientific data will require standardization and sustained commitment.

Life science companies should look to software solutions that embed semantic enrichment to find relevant scientific concepts faster and to accelerate new discoveries.

> There remains a need for a search tool capable of leveraging evidence based biological connections to show researchers datasets useful for hypothesis generation or scientific support.
>
> Waldrop et al

Applying data analytics

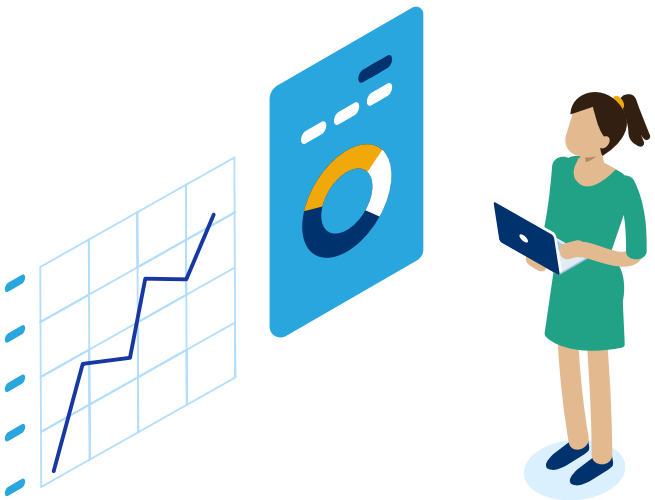
The use of analytics to drive decision making in life sciences is nothing new. Yet, as the explosion in data and information has spread pervasively across research and development organizations, putting these treasure troves of data to use has become a priority. Novartis, for instance, has spent the last several years developing a state-of-the-art advanced analytics platform centered on data science, called Nerve Live (Finelli & Narasimhan, 2020). Its goal is to embed insights-driven decision making across its drug development pipeline and to help the clinical development team access and apply trusted, relevant information more easily.

Where it comes into play

We've discussed how applying search, personalization, and semantic enrichment helps to narrow down the vast amounts of literature and data at our disposal. Analytics and visualization technologies — such as knowledge graphing, clustering, and ranking — are additional ways to help weed out unnecessary information and give users additional ways to explore the information and the connections between disparate sources.

It takes five minutes, on average, to read one page of technical material (Speed Reading Facts, n.d.). Consider the impact on researchers who has one fewer scientific paper to read every day because data analytics and visualizations were applied to allow them to focus on key pieces of information and be more selective about which papers require a full read.

Knowledge graphs show secondary datasets created by analyzing and organizing underlying data. The visual element of "seeing" new connections and naming key relationships can accelerate the generation of novel insights. For example, analytics embedded within a company's information management system could provide capabilities such as topic trending or correlating topics related to a search term. While such analytics provide time savings for researchers they also have the potential to identify previously unassociated connections and support researchers in uncovering novel insights.



Challenges and opportunities

The growing field of Literature-Based Discovery (LBD) shows us the potential for data analytics. Don R. Swanson was the first information scientist who searched scientific literature to establish a previously unknown link between a disease and a potential treatment. In a mostly manual fashion, he analyzed publicly available scientific literature related to fish oil and Raynaud's Disease, and found that the concepts of blood viscosity, platelet aggregation, and vascular reactivity were implicitly shared yet never explicitly connected (Swanson, 1986). Subsequent clinical trials proved out Swanson's hypothesis that fish oil could successfully treat Raynaud's Disease, and his seminal paper published in 1986 launched LBD as a new research field.

Though easier said than done, the ability to extract knowledge from data remains a primary goal and a key competitive advantage for life science companies. Organizations should consider embedding data analytics and visualization as a strategic capability. This would require investing in data scientists, establishing new ways of working across functions, developing data literacy through training and just-in-time opportunities, and deploying data visualization tools.

The ultimate value and utility of data analytics and visualization, such as knowledge graphing, clustering, and ranking, lies in their power to unearth vast amounts of undiscovered knowledge more quickly and efficiently. Powerful analytics in information and knowledge management solutions — buoyed by the right people — not only forge but can accelerate new connections between medical concepts residing in scientific literature and data.

Conclusion

Enabling scientific content discovery and managing disparate data will only continue to become more complex due to the ever-increasing volume, velocity, and variety of newly generated information. Effective enterprise solutions are available to help manage the "information chaos" hampering the work of functions such as R&D, Regulatory, Safety, Scientific and Medical Affairs, and Competitive Intelligence. However, organizational leadership is required to elevate the strategic importance and awareness of intelligent information management and to invest in the technology, people, and processes that will drive novel discoveries and insights.

Realizing the full potential of streamlined search, usage, and insight generation requires:

1. automated integration of all relevant external information with internal intellectual property,
2. elimination of data silos,
3. connection between concepts, and
4. organizational development such as hiring new talent, embedding strategic capabilities, aligning leadership, and implementing new processes and tools.

Those companies that "learn" more quickly than others are in a better position to "fail fast," adapt, pivot, and focus on the optimum opportunities, with a higher likelihood of success. While streamlining the information retrieval and "sense-making" process can accelerate research and development, it would also likely **create a competitive advantage**. Finding the right software and an expert technology partner can help design and implement targeted solutions.

Learn more about RightFind

RightFind Navigate provides personalized search across multiple sources of data and information for highly relevant discovery. RightFind Navigate is a part of the RightFind Suite, a robust set of software solutions that fuel scientific research and simplify copyright, anytime, anywhere.

About Imbue Partners

Imbue Partners is a specialized consultancy that partners with biopharma and medical technology organizations around the world to co-create and implement tailored strategies and solutions in customer centricity, market growth, product launch, and competitive readiness. Imbue collaborates with commercial, medical, and patient experience/engagement teams to identify opportunities, boost capabilities, and eliminate roadblocks. In doing so, these teams can achieve sustainable, positive change for their patients, their people, and their business. Founded in 2009, Imbue Partners is a certified Women Business Enterprise (WBE) headquartered in Middleton, Massachusetts.

Acknowledgements

This work was supported by underwriting from Copyright Clearance Center.

Considerations

In this paper, we have generally adhered to the seventh edition of the American Psychological Association (APA) formatting guide.

References

1. *1 Second - Internet Live Stats.* (n.d.). https://www.internetlivestats.com/one-second/
2. AIIM. (2021). AIIM 2021 State of the Intelligent Information Management Industry: A Wake-Up Call for Organization Leaders. In http://www.aiim.org/research. Retrieved December 9, 2022, from http://www.aiim.org/research
3. Bohannon, J. The Secret Half-Lives of Scientific Papers, Science, December 19, 2013, https://www.science.org/content/article/secret-half-lives-scientific-papers
4. Dean, B. (2020, August 20). How People Use Google Search. Backlinko. Retrieved December 9, 2022, from https://backlinko.com/google-user-behavior
5. Finelli, L. A., & Narasimhan, V. (2020). Leading a Digital Transformation in the Pharmaceutical Industry: Reimagining the Way We Work in Global Drug Development. Clinical Pharmacology &Amp; Therapeutics, 108(4), 756–761. https://doi.org/10.1002/cpt.1850
6. Karamanou, M., Protogerou, A., Tsoucalas, G., Androutsos, G. & Poulakou-Rebelakou, E. Milestones in the history of diabetes mellitus: The main contributors, World Journal of Diabetes, January 10, 2016, https://doi:10.4239/wjd.v7.i1.1
7. MeSH Browser. (n.d.). https://meshb.nlm.nih.gov/
8. *Speed Reading Facts.* (n.d.). https://www.execuread.com/facts.htm
9. Swanson, D. R. (1986). Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1), 7–18. https://doi.org/10.1353/pbm.1986.0087
10. Waldrop, A. M., Cheadle, J. B., Bradford, K., Preiss, A., Chew, R., Holt, J. R., Kebede, Y., Braswell, N., Watson, M., Hench, V., Crerar, A., Ball, C. M., Schreep, C., Linebaugh, P. J., Hiles, H., Boyles, R., Bizon, C., Krishnamurthy, A., & Cox, S. (2022). Dug: a semantic search engine leveraging peer-reviewed knowledge to query biomedical data repositories. *Bioinformatics*, 38(12), 3252–3258. https://doi.org/10.1093/bioinformatics/btac284
11. Ward, A.F. (2013). *One with the Cloud: Why People Mistake the Internet's Knowledge for Their Own* [Doctoral dissertation]. Harvard University.

Copyright Clearance Center (CCC)

A pioneer in voluntary collective licensing, Copyright Clearance Center (CCC) helps organizations integrate, access, and share information through licensing, content, software, and professional services. With expertise in copyright and information management, CCC and its subsidiary RightsDirect collaborate with stakeholders to design and deliver innovative information solutions that power decision-making by helping people integrate and navigate data sources and content assets.

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

U.S. organizations:

copyright.com/rightfind

solutions@copyright.com

Outside U.S. organizations:

rightsdirect.com/license/rightfind-enterprise

solutions@rightsdirect.com