Copyright Clearance Center | RightsDirect

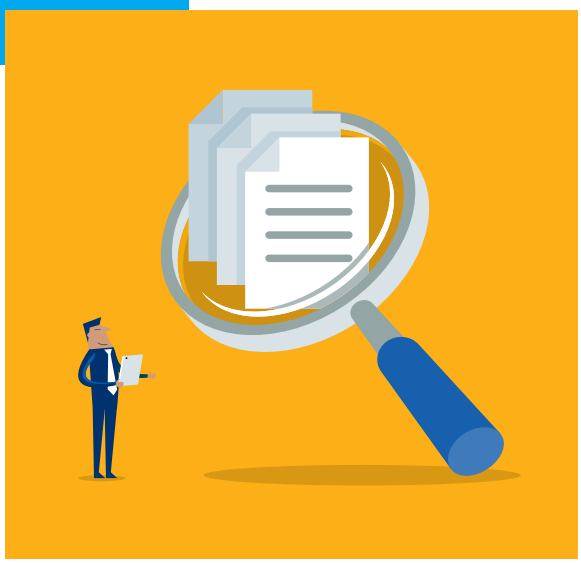
Whitepaper

# Benefits of Semantic Enrichment Across the Drug Development Pipeline

Semantics is the process of assigning meaning to words and is a vital step in making unstructured scientific content machine readable, contextualizing otherwise unstructured information. It unlocks a host of benefits to search and analytics which are outlined below in the context of use cases relevant to drug discovery.

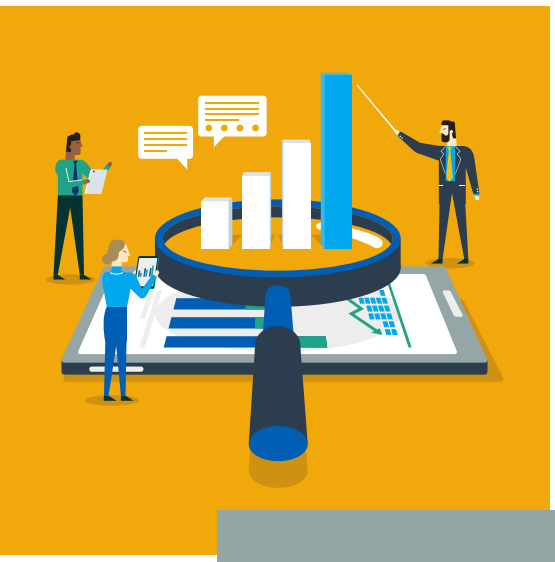
Applying semantic enrichment within an information system enables researchers to search and analyze entities with real meaning. Linking entities is the next step to reduce the time it takes to digest a large, unstructured and heterogenous content set. Thanks to semantic technology, researchers can identify the nuggets of information that make a difference in a matter of minutes, instead of the hundreds of hours that it would take to do the same analysis manually.

Practical applications of semantic search across the drug development pipeline enable life sciences organizations to not only save time but increase the accuracy and efficiency of their processes. This white paper discusses four practical use cases using this technology with various stages of drug development to show the benefits that leveraging semantic search can bring to the table.



Use Case #1

# Competitive Intelligence



Pharmaceutical companies must keep current with the latest biomedical research relevant to their own therapeutic programs. However, the accelerating amount of data generation and increasingly diverse range of sources make it challenging to maintain a comprehensive and up-to-date understanding. To ensure competitiveness, they must move on from the prevailing manual approach involving the time consuming, piecemeal review of a small range of data sources. But automating the process of scanning multiple sources of data is challenging because terms or phrases of interest can be spread out in an article and different authors use different terminology when describing the same thing.

Most literature searches are typically limited to finding the specific terms used by the author. Through semantic enrichment, however, all relevant data is found, regardless of which synonym is used as the search term. Established controlled vocabularies which apply an explicit, unique meaning and description to scientific terms provide comprehensive coverage of relevant terminology and the robust foundation necessary for an effective and impactful literature monitoring strategy.[1]

## 2 Questions Every Competitive Intelligence Team Should Ask

1 **Do we have a way to monitor competitive information in a collaborative way?**

When articles, press releases, journals and other data are stored in different systems and locations, you cannot be sure everyone charged with monitoring competitive intel has access and rights to reuse that content. When employees store content on their individual hard drives, for example, it is difficult to collaborate on new materials coming in.

Using one convenient workflow to purchase, organize, and collaborate on content can provide competitive intelligence teams the tool they need to collaborate on information. With a shared library tool, users can create libraries for approved team members where all competitive information related to a specific product, therapeutic, or topic of research can not only be accessed, but also annotated and highlighted. Automated alerts can also be set up to help individuals or teams receive the most relevant information about their focus area without having to search for it.

2 **Do we have a platform to search across data sources in aggregate?**

The ability to access content and data such as clinical trial information, drug and device patents, sales forecasts, scholarly articles, preprints, and more, is critical to gain a well-rounded perspective on the market.

When employees must consult multiple sources to find all the content and data they need, it's a massive manual undertaking. By bringing together content and data from publicly available sources along with licensed content and internal proprietary data in a single interface, connecting the dots within competitive intelligence information becomes an intuitive process instead of a manual one.

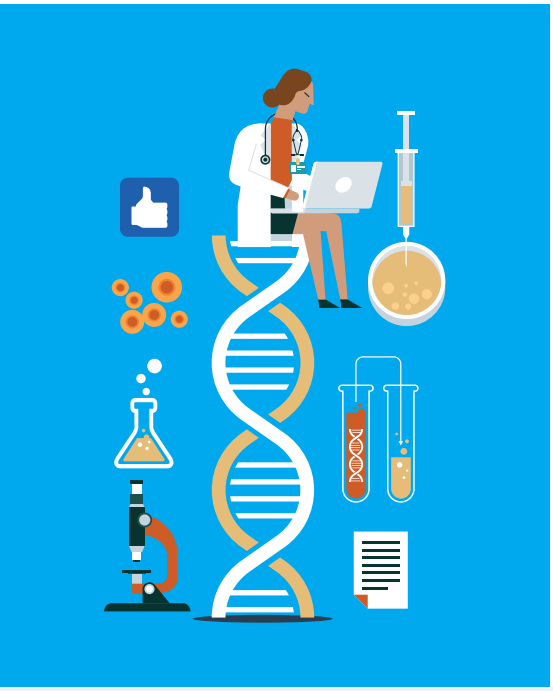
Use Case #2

# Drug Safety

In most countries, adverse event reporting is now a regulatory requirement for pharmaceutical companies. Pharmacovigilance and drug safety teams are challenged to maintain safety and compliance with the same, or fewer, resources amid increasingly stringent, globally diverse regulations.

Traditionally, many pharmaceutical companies have been conservative in incorporating machine analysis technology into a pharmacovigilance workflow, largely due to compliance concerns. However, with the amount of data and the number of sources growing exponentially, companies have started to test and integrate machine analysis tools into their pharmacovigilance workflows, as the process of monitoring, identifying, and analyzing adverse events across all these sources can be quite manually intensive.

A pharmacovigilance workflow typically involves many different data sources, including patient cases, healthcare reports, scientific literature, and even social media.



## 3 Ways Pharmacovigilance Teams Can Utilize Semantic Search[2]

- **Literature Review:** Semantic search enables articles to be ranked by their relevance to a particular search and visually highlights all relevant terms and concepts present within each article, without being limited by the indexing terms used by the data source.
- **Automated Surveillance:** Semantic search can simplify the process of flagging safety signals and continually scanning all available sources (think "pre-screening" in near real-time), and only issues alerts when something noteworthy is found. This can be applied to any process whereby prioritization and/or routing of documents can only be determined by first reviewing the content of the document.
- **Predictive Analytics:** Pharmacovigilance teams can utilize a solid bedrock of semantically enriched content to be automatically alerted to interesting new connections as they are found. Since any new knowledge can be used to refine the surveillance process, a feedback loop can be established to continuously improve and expand organizational knowledge.

A pharmacovigilance workflow typically involves many different data sources, including patient cases, healthcare reports, scientific literature, and even social media.

Use Case #3

# Artificial Intelligence/Machine Learning[3]

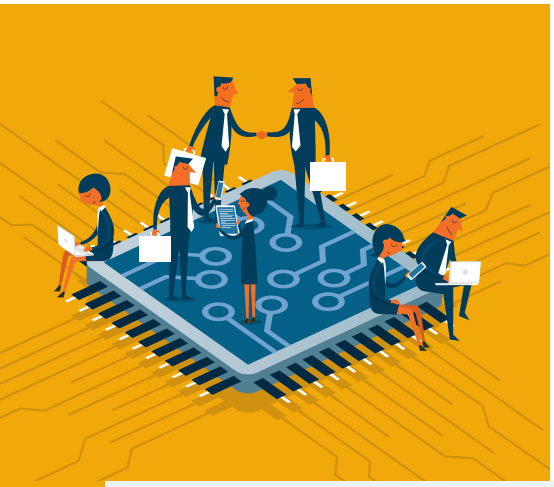
The combination of Artificial Intelligence (AI) and big data is triggering a revolution across the entire drug development lifecycle; from the way new drugs and treatments are discovered, to identifying opportunities to re-purpose those already in the market.

Within the pharmaceutical and healthcare sectors, big data represents a great hurdle as approximately 80% of clinical data is stored as unstructured text. AI techniques are therefore required to identify concepts, entities and relationships within the document corpus. While the volume and variety of big data represents a major technical challenge to any pharmaceutical organization, the payoffs are also substantial, enabling patterns and trends to be identified that can inform decision making at all stages of the drug development process.

By combining AI and deep learning models with semantic algorithms, organizations can exploit data and accelerate its downstream use in research and development.

## 3 Ways the Unique and Powerful Combination of Semantic Technologies and Deep Learning Can Train Models

- **Named Entity Recognition (NER):** Help ensure all references to entities of interest and their synonyms are identified accurately, regardless of whether they are present in an ontology or not. By identifying potential new entities and synonyms, semantic deep learning models augment and extend existing ontologies and rapidly develop ontologies for new domains, which can in turn be used to further train the AI.
- **Predictors:** Spot patterns in data that help predict future outcomes.
- **Clustering and classification:** Group documents and concepts based on their underlying data relationships through clustering or classification.



In order to train such models, however, you need large volumes of content and semantic enrichment to harmonize and clean that content in order to create high-quality training data. In the case of pharmacovigilance, for example, the combination of semantic enrichment and deep learning can help pharmaceutical companies avoid the trade-off between the volume of content and the time available to review it, as well as identify and prioritize potential adverse events from the growing corpus of scientific literature with high confidence.

## Breaking Down Unstructured Text in External and Internal Data

Today, 80% or more of an organization's data is held in unstructured text such as Word documents, PowerPoint slides and PDFs.

### But this is also true of external data sources such as:

Patents | Blogs | Drugs | Conference materials | Clinical notes | Preprints | Literature databases

Use Case #4

## Knowledge Graphs[4]

Knowledge graphs depend on the ability to map between equivalent concepts from many heterogeneous data silos – internal or external. Semantic annotation tools scan vast quantities of scientific data sources and normalize these scientific concepts to unique entity IDs. These IDs are then mapped to a whole host of public standards, to pull together evidence mined from the literature, alongside supporting evidence from the public domain or an internal data system.

Knowledge graphs capture relationships between entities and make it easier to index, process, and find "nuggets of knowledge." They help us think in a reference frame that grounds our interpretation of data and enables us to search for knowledge by using "things" rather than "strings."

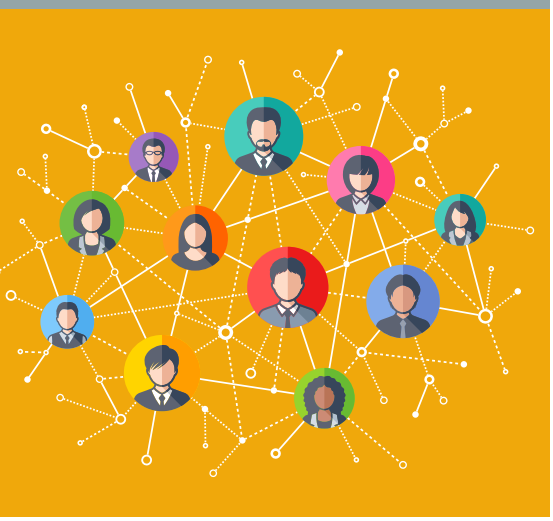
Imagine a research team is working on a new drug that might be effective against a particular disease. A simple literature search might find multiple instances of the indication or even the biological target mentioned alongside a particular molecule. However, we often need to go deeper than that and explore potential side effects, predict what the best mechanism for administration might be, or find ways to avoid issues through medicinal chemistry and molecule design. This is where a knowledge graph can help researchers make those leaps of understanding by connecting apparently disparate pieces of information.[5]

Investment in knowledge graphs pays off because they are not single-purpose models; they inform multiple decisions, not just one. Knowledge graphs are like living organisms, they are constantly changing, and we can continuously benefit from our investments as a result.

When we use knowledge graphs in combination with technologies such as machine learning, our data output moves up from simply being "stated" to being "informed." This gives us greater confidence in the inferences we make from the data and provides us a better ROI from our investments in graphs.

### Licensing Considerations for Knowledge Graphs

The semantic connections and predictive analytics surfaced through knowledge graphs can be extremely valuable, but to ensure users' trust in these sources, they always need to be validated through the review of the underlying research. That means the ideal approach is a workflow that highlights the key information in the form of a knowledge graph but allows for easy access and use of the full-text content for deeper insights. And because many use cases depend on more than one individual, proper licensing must also be in place to support collaboration. Licenses are typically also required to process content to build machine learning models and extract insights through knowledge graphs.



## 2 Ways to Approach Knowledge Graphs

There are typically two approaches to creating knowledge graphs:

- **Enterprise Level:** An enterprise knowledge graph will be more abstract, and include data from many departments within an organization, for example: Finance, HR, Legal, and R&D. In these circumstances, users are viewing the data from different aspects or through a particular lens.
- **Project Level:** At a project level, the use case is more clearly defined: what specific questions do we want to ask of the knowledge graph? For example, an exercise in target prioritization will require focus on gene-disease associations and relevant datasets.

SciBite
an ELSEVIER company

SciBite's data-first, semantic analytics software is for those who want to innovate and get more from their data. Leading the way by pioneering the combination of the latest in machine learning with an ontology-led approach, SciBite's semantic infrastructure answers business-critical questions in real time by releasing the value and full potential of unstructured data. Supporting the world's leading scientific organisations with use-cases from discovery through to development, SciBite's suite of fast, flexible, deployable API technologies empower customers, making it a critical component in scientific, data-led strategies. SciBite is headquartered in the UK with additional sites in the US and Japan. Find out more at www.scibite.com

# Conclusion

CCC is pleased to offer **RightFind Navigate with Semantic Search**, a solution that helps to identify hidden connections between data, breaks down silos, and gets researchers to the content they need, faster. RightFind Navigate with Semantic Search leverages SciBite biomedical vocabularies to identify entities across large volumes of unstructured content including scientific literature, preprints, drugs, clinical trials, and global life science patents, among others.

*SciBite extracts meaning from unstructured data via rules-based named entity recognition (NER) using expertly curated ontologies and updated by leveraging the latest machine learning technologies. SciBite's NER-tuned vocabularies cover dozens of scientific domains and allow for high-speed tagging of relevant, disambiguated terms from your documents.*

## Resources

[1] See SciBite Use Case: Comprehensive competitive intelligence monitoring in real time.
[2] See SciBite Use Case: Semantic Analytics: An Integrated Approach for Pharmacovigilance Teams to Achieve Total Awareness
[3] See SciBiteAI in depth
[4] See SciBite Solutions: Knowledge Graphs
[5] See CCC White Paper: Knowledge Graphs

**About CCC**
A pioneer in voluntary collective licensing, CCC (Copyright Clearance Center) 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.



CCC | RightsDirect

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

U.S. organizations:
copyright.com/rightfind-navigate
solutions@copyright.com

Outside U.S. organizations:
rightsdirect.com/rightfind-navigate
solutions@rightsdirect.com