Use case

Unlock the full potential of departmental scientific documents
Unlock the full potential of departmental scientific documents

Much of the knowledge found within Pharmaceutical departments or cross-functional project teams is contained in the plethora of documents, reports and emails that they produce and the scientific articles they have downloaded. The unstructured nature of these files, the range of formats used and the fact that they are typically spread across different locations limits the ability to mine them for useful information.

Even where such documents are organised in a file store or structured in a document management system, the accompanying search capabilities are limited to exact matches of what was written by the document author. Similarly, inconsistent use of synonyms during data entry makes it difficult to identify and collate all relevant data for a disease or target of interest. For example, a document describing work on ‘muscarinic acetylcholine receptor M1’ will not be found by anyone searching for the commonly used synonym ‘cholinergic receptor muscarinic 1’.

SciBite provides scientific teams with the opportunity to semantically enrich their documents, opening up new possibilities to mine the data more effectively and derive valuable insights.

SciBite's ontology management platform, CENtree, enables organisations to maintain up to date ontologies representing evolving scientific language. Subject matter experts can easily contribute to keeping things current and augment our manually curated standard reference vocabularies (such as Cell, Gene and Indication) with proprietary information, such as project codes and IDs used to track materials such as compounds and cell lines, to create a single, authoritative, universally applied terminology. CENtree also leverages machine learning techniques to suggest potential new terms that are similar to other related words because they are used in a similar context.

When used in combination with CENtree, SciBite's Named Entity Recognition (NER) engine, TERMite, applies standard, well established ontologies and controlled vocabularies to scientific documents, generating a semantic index and endowing text entries with an explicit, specific meaning. This transforms unstructured document text into a structure that can be queried in a simple fashion to answer questions that would otherwise require time-consuming, error-prone manual work. As illustrated below, the semantically enriched data can be surfaced through our easy to use SciBite Search interface, via 3rd party search and visualization tools such as Spotfire and Linkurious, or a combination of these.
The benefits of semantically-enriched data

**Smart, comprehensive searches**

Most documents repositories only have rudimentary search capabilities. For example, a search of a typical document store for the Alzheimer’s related gene, PSEN1, would miss references to synonyms such as Presenilin-1, AD3 and PSNL1.

**Using a rudimentary text query, a search for the gene ‘PSEN1’ would find this….**

The gene **PSEN1** was amplified by PCR using the primers listed below….  

**…but would miss all of these documents referencing synonyms of PSEN1…**

- An assessment of the effect of a series of **Presenilin-1** mutations…
- Genetic characterization of the early-onset Alzheimer’s disease **AD3** locus
- PSNL1 gene expression in human tissues at the tissue and cellular level

**Figure 2: Rudimentary text searches miss synonyms of search terms of interest**

Through semantic enrichment, SciBite ensures that all relevant data is found, regardless of which synonym is used as the search term. SciBite not only makes it simpler to interrogate the information managed within internal documents, it also enables users to answer more complex ontology-based questions and find all documents related to a specific topic e.g. all those that talk about mode of action, all those that are about toxicology, etc.
Identify important themes
SciBite accurately marks-up all relevant terms and concepts within a document, enabling scientists to rapidly identify the topics covered in an experiment, easily interpret the text and instantly get a feel for what a document is about.

Examples of questions that can be rapidly answered with semantically enriched document data

- Find all references to project ABC-101, regardless of the syntax used by the author (e.g. ABC-101, ABC101 and ABC 101)
- Find all experiments that reference a compound of interest used in combination with one or more other compounds of interest
- Find all experiments for a specific target across the organisation, regardless of which synonym was used by the author of the experiment
- Which projects are focussed on biologics?
- Which targets have we studied that are associated with inflammatory disorders?
- Which diseases have we studied for both a target of interest and other targets in the same class?
- Which pre-clinical studies have utilised a specified mouse model?
- Which experimental techniques are growing across the organisation and would benefit from a core facility?

Figure 3: Summarising all terms identified within a document (right-hand side of the screen)
SciBite enables users to perform searches for terms that co-occur within a sentence or within a document. For example, by generating a list of genes which are mentioned most frequently with a disease of interest, potential new avenues for research can be revealed.
**Figure 4:** A ranking of the terms that co-occur most often with a target of interest
Explore Connections

When different authors use different terms to describe the same thing, such as ‘heart attack’ and ‘myocardial infarction’, it can be difficult to make connections between them. SciBite provides the “semantic glue” to enable cross-linking between equivalent concepts described by different documents, revealing otherwise hidden associations.

SciBite can also make connections by identifying significant terms mentioned in a document and presenting users with a ‘More like this’ list of other documents that include similar topics. SciBite can also reveal associations between individual scientists and the topics mentioned in their documents, which helps identify experts in a particular disease or methodology, regardless of what is written in their profile.

Figure 5: Interactive relationship network of documents. Individual documents are shown as green circles. Other circles indicate semantic annotations shared with one or more other documents, highlighting relationships between these. Visualisation is of a neo4j database using the Linkurious browser tool.
Gain a holistic view
In addition to document data, SciBite can also apply semantic enrichment to other internal sources such as electronic lab notebooks, publicly available literature (such as MEDLINE) and databases (such as OMIM and Orphanet), patents and clinical trials data (such as ClinicalTrials.gov) and third-party data, such as commercially available databases. Since the resulting data will be as well-structured and interoperable as public data, it becomes facile to integrate multiple disparate sources, search across internal and third party data in a single interface and easily gain a holistic view of the knowledge associated with a compound, target or disease of interest.

Users can understand the trends associated with any topic of interest and ask questions across data sources that would have otherwise been time consuming or impossible to answer. For example, it becomes easy for a department to define its strengths and understand its competitive position relating to specific targets or diseases and assess if it is working on things that its competitors aren’t (and vice versa). Similarly, it is possible to identify which companies or institutions are working in which disease or technology area of interest to explore options for collaboration. In each case, alerts can be setup to ensure information is highlighted to the right people in a timely manner.

Complementing internal search tools
All the core functions of the SciBite platform can be accessed programmatically through its rich powerful RESTful application programmatic interface (API). For example, SciBite enables any web-based search interface to become semantically intelligent. Instead of being presented with restrictive and lengthy drop-down menus, users can enter text into semantically aware search fields and have relevant terms suggested to them as they type. Instead of being presented with restrictive and lengthy drop-down menus, users can enter text into semantically aware fields and have relevant terms suggested to them as they type.

Figure 6: Semantically-intelligent data entry forms
Summary
Most departments and teams in pharmaceutical companies are unable to realise the true value of the data stored in the documents they produce. SciBite enables scientific documents to be searched intelligently, facilitates integration of the data within them with other sources and makes them computationally accessible for automated analysis.

SciBite delivers a cost-effective solution to unlock the wealth of information managed within departmental documents, opening up new possibilities to mine the data more effectively and derive deep scientific insights.

About SciBite
SciBite is an award-winning semantic software company offering an ontology-led approach to transforming unstructured content into machine-readable clean data. Supporting the top 20 pharma with use cases across life sciences, SciBite empowers customers with a suite of fast, flexible, deployable API technologies, making it a critical component in scientific data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com
SciBite’s data-first, semantic analytics software is for those who want to innovate and get more from their data. At SciBite we believe data fuels discovery and we are leading the way with our pioneering infrastructure that combines the latest in machine learning with an ontology-led approach to unlock the value of scientific content. 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 our customers, making it a critical component in scientific, data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com