

Use case

SciBite: Identifying internal expertise



SciBite: Identifying internal expertise

Have you ever struggled to find people in your organisation with expertise that could help your research project? It's a common problem, particularly for global pharmaceutical companies with geographically dispersed R&D teams. Having an understanding of the availability and location of internal expertise is fundamental to effective and efficient project planning. Building new cross-disciplinary teams, identifying stakeholders for strategic initiatives, and evaluating potential sites for new centres of excellence all benefit from having accurate information about an organization's core competencies and expertise.

A person's expertise is often reflected in the content they generate, such as reports, presentations, experimental write-ups, emails and posts to internal messaging boards. While these sources contain valuable data, there are several barriers to extracting it and using it effectively. Firstly, the unstructured nature of these sources, the variety of formats and locations and the data privacy conditions of each source all limit the ability to mine them for useful information.

Even where documents are organised to some degree in a repository, searches are typically limited to keywords. 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'. Similarly, the use of informal

vernacular, such as the range of different abbreviations and 'nicknames' people use to describe projects, present a challenge to most computational approaches.

Finally, simply because a scientist mentions a disease or target of interest, one shouldn't infer that they are an expert - there is an inherent high level of 'noise' in internal data sources that hinders the identification of mentions that are actually relevant.

SciBite enables pharmaceutical companies to index all the content generated by a scientist and build a profile of relevant terms that reflect the research activities of that person. These knowledge profiles can uncover valuable insights about the latent expertise that resides within an organisation.



Figure 1: Semantic enrichment of unstructured data sources



Annotating scientific terms and author information in unstructured content

With the appropriate data privacy policies in place, SciBite enables customers to ingest a wide range of file types, including emails, Word documents, PowerPoint presentations, CSV files, PDFs and ELN records¹, and transform the unstructured text within them into richly annotated machine-readable data.

SciBite uses established controlled vocabularies, or VOCabs, to apply an explicit, unique meaning and description to scientific terms contained within a document.

This enables scientific text to be contextualised and machine-readable data, irrespective of its source.

Through semantic enrichment, SciBite ensures that all relevant data is found, regardless of which synonym is used as the search term. For example, a search for the Alzheimer's related gene, PSEN1, will also find mentions of Presenilin-1, AD3 or PSNL1.

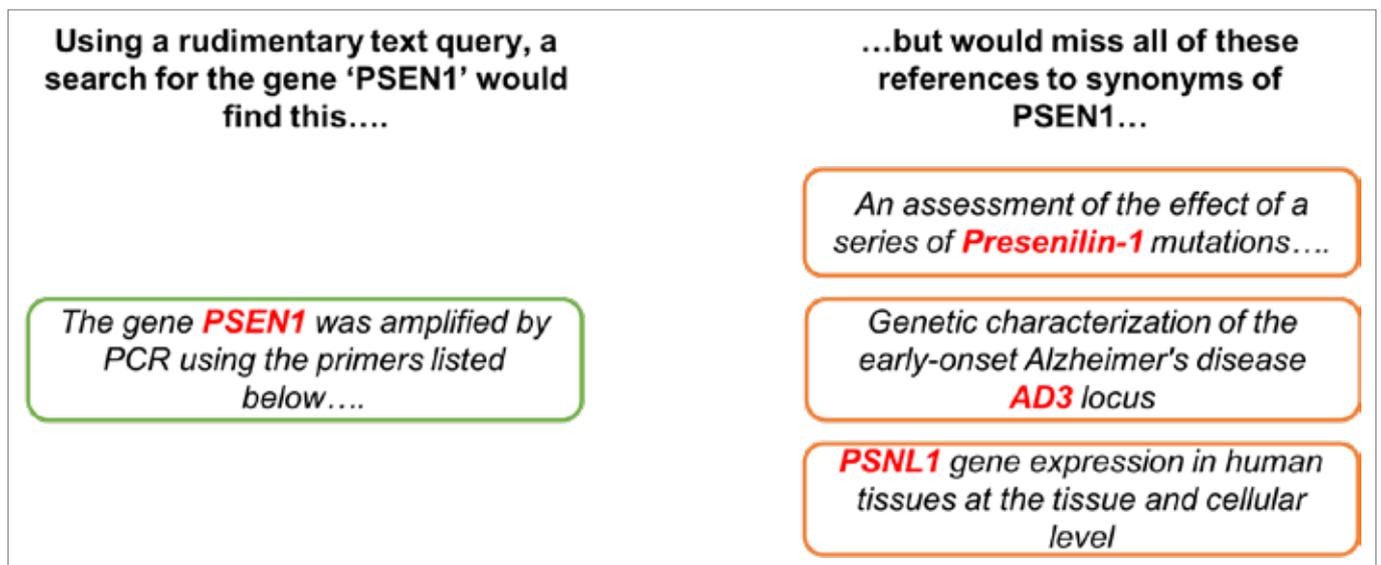
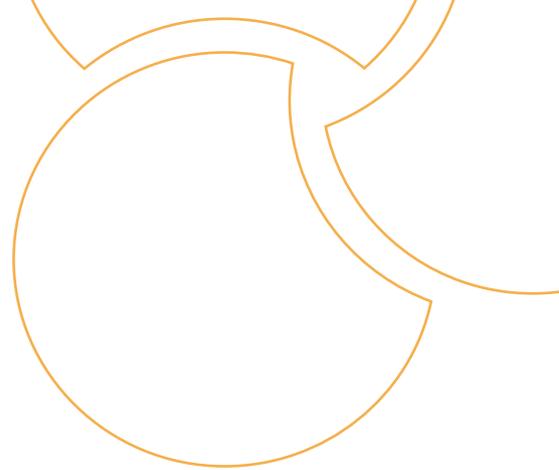


Figure 2: Keyword and text searches miss synonyms of search terms of interest

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 departmental names, project codes, cell line IDs and internal drug abbreviations. Similarly, SciBite can transform a list of employees into a vocabulary. For example, an Active Directory list can uniquely link an individual to all internal

content that they author, even if multiple people within the company share the same name.

By leveraging internal information, SciBite can annotate documents with author metadata and connect this with both public and proprietary terms mentioned in the documents they write. This enables users to ask questions using both internal and external terms, such as finding which people are connected to both a disease and a drug candidate of interest.



Ensuring relevance

Not all entities mentioned within a document are equal. For example, the text below is about IL-6, but several other genes and diseases are also mentioned, making it challenging to rapidly identify what the document is actually about.

Similarly, a scientist who mentions a gene in a report isn't necessarily an expert on that gene, whereas someone who does so tens or hundreds of times probably is.

SciBite addresses this issue by applying quantitative, qualitative and linguistic rules to rank mentions of terms within a document based on their significance. The combination of the frequency at which a given author mentions a term and the significance of each mention provides increased confidence that a person has relevant expertise related to that term. By filtering through the noise, the most relevant terms can be assigned to a person's knowledge profile.

“Overexpression of IL-6 is associated with a number of diseases including multiple myeloma, rheumatoid arthritis and psoriasis. IL-6 is a member of the family of cytokines including interleukin-11, leukemia inhibitory factor, ciliary neurotrophic factor and oncostatin M, cardiotrophin-1.”

Figure 3: Extract of a document which references multiple genes and diseases

Mining expertise

Individual knowledge profiles can be aggregated to generate a searchable capability map of the organisation. For example, if a Chief Scientific Officer reads a news article about a novel technology, such as CRISPR, they can rapidly determine if scientists in their own organisation are utilising it and, if so, who has demonstrable knowledge about that topic or combination of topics.



Figure 4: While a Sue may be better known for her work on PPARG, her knowledge profile reveals a high frequency of relevant mentions of cyclin-dependent kinase Inhibitor 2A, CDKN2A

Users can proactively search for colleagues across the organisation with overlapping knowledge profiles and identify opportunities to collaborate or build a new team.

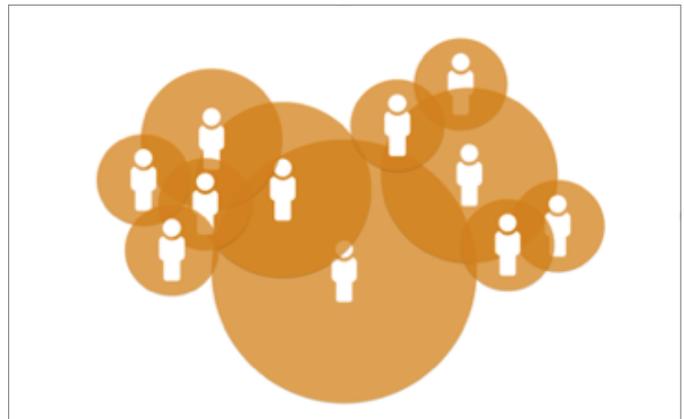


Figure 5: Identifying people with similar or overlapping knowledge profiles



Similarly, it is easy to pinpoint hyper-connected individuals in the organisation that can help improve communication between departments or sites.

SciBite can also reveal the trending topics within an organisation and enable users to quickly identify which scientists are at the forefront of research in each of these areas. At a higher level, it becomes more straightforward to define the research profile of a business unit or the wider organisation, identify potential resource gaps and evaluate the relevance of new partnerships or acquisitions.

Summary

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

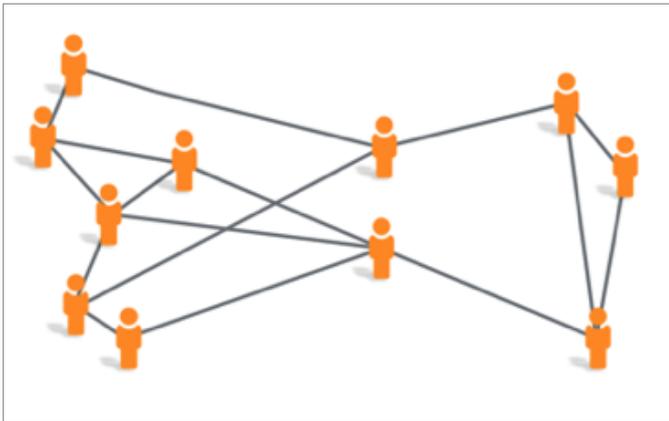


Figure 6: By aggregating users knowledge profiles, it is possible to identify hyper-connected individuals

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

Head Office:

SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

 www.scibite.com
 contact@scibite.com
 LinkedIn: SciBite
 Twitter: @SciBite
 +44 (0)1223 786 129

