

Use case

Enabling better data discoverability: Leveraging semantics for intelligent scientific search





Enabling better data discoverability:

Leveraging semantics for intelligent scientific search

SciBite Search delivers intelligent scientific search to everyone, you don't have to be a search expert. Its modern, easy to use interface provides scientists with access to domain specific ontology and AI-powered search capabilities to quickly find the answers they need while also providing sophisticated text mining capabilities for deeper and more exploratory queries.

SciBite Search enables users to search both unstructured and structured information across:

- Public biomedical sources, such as Medline and Clinical Trials.gov,
- Subscription content from providers such as Elsevier, Wiley and Springer, and
- Internal and external documents.

Document-level security and role-based permissions ensure that access to proprietary information is controlled.

Here we highlight just a selection of the many applications of SciBite Search.

Intelligent scientific search, available to everyone

Most search applications are limited to rudimentary text and keyword searches and lack the scientific intelligence needed to answer common research questions. For example, a search for the gene PDE5A would miss references to synonyms such as CGB-PDE, Phosphodiesterase 5A and CN5A.







At the core of SciBite Search are our scientific vocabularies, or VOCabs, and Named Entity Recognition (NER) engine, <u>TERMite</u>. SciBite's team of experienced ontologists create and maintain VOCabs by augmenting public ontologies using a combination of manual and machine curation to ensure quality and accuracy and ensure they have comprehensive synonym coverage. Using these VOCabs, TERMite rapidly processes all ingested content and applies an explicit, unique meaning and description to scientific terms and concepts found within the text, contextualizing them so that they can be understood as "things, not strings"1. For example, SciBite Search understands that PDE5A is a protein rather than just a collection of characters. This ensures that all relevant data will be found, regardless of which synonym is used as the search term.

This contextualisation also enables SciBite Search to answer semantic questions. For example, an article describing "The Effect of Sildenafil on Selenite-Induced Cataract in Rats" would be found by a simple text search using any of the terms in the sentence. However, SciBite Search understands, for example, that Sildenafil is a drug which inhibits the target PDE5A and that rats are rodents. So, a query for 'phosphodiesterase inhibitors in rodents' would return exactly the same article.





¹ https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html



SciBite Search also enables scientists to use natural language scientific queries, surfacing powerful query capabilities while hiding the complex logic from end users. This ranges from questions that have a 'yes/no' answer to those that generate set of facts and those that require aggregation of information, to the such as:

- Is Sildenafil an inhibitor of Phosphodiesterase 5A?
- What are the known inhibitors of Phosphodiesterase 5A?
- What is the top drug for Phosphodiesterase 5A?
- What are the top 5 journals for cancer research?

op 5 journals for cancer	¥ - Q [1	DOCUMENTS	SENTENCES	•
MEDLINE 2,189,075 JOURNAL 13,071 TRIAL 4,028 PDF 7					
lowing 1 - 20 out of 284,580		RESULTS PER PA	GE: 20 ¥	SORT: RELEVAN	CE
What are the top 5 journals for Neoplastic Cell Transformation	n				
The top 5 journals for Neoplastic Cell Transformation in documents					
annals of oncology : official journal of the european society for medical oncology				12,3	77
scientific reports				4,3	172
alos one				3,6	689
european radiology				3,0	95
					18

Figure 3: SciBite Search enables users to easily answer complex questions

Deeper search and exploration

As well as improving the search experience for scientists, SciBite Search also provides powerful text mining capabilities to support more sophisticated use cases.

Document search tools are limited to the content of the document index. For example, such tools will only know that Viagra is made by Pfizer if this relationship is explicitly stated within a document. SciBite Search leverages knowledge graphs and deep learning to augment the scientific search experience, enabling users to make connections between scientific concepts and find information based on relationships. For example, users can ask questions that rely on the metadata of the entities mentioned in documents, such as looking for "pharmaceutical companies in the south of the United Kingdom mentioned in the same sentence as cancer".

Users can perform boolean queries and quickly navigate to the most interesting results based on taxonomies. SciBite Search also enables users to instantly identify 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.



Chrome File Edit View History	/ Bookmarks Peopl	e Tab Window Help	😘 💽 🕄 🛜 🌒 100% 🚱 Thu 10 Sep 17:46	s phil verdemato 🔍 🗮
🔍 🔍 🎓 SciBite Search Sentences 🛛 🗙	DocStore Swagge	r API - Custe 🗙 🔸		
> C docstore2.scibite.io/sente	ences?document_type	=medline&offset=40&queryString=pde5a%20or%20heart		x 0 * 0 :
🗄 Apps 👩 Phil Verdemato 🏹 Base CR	RM 🥝 Postman 🚱 i	nteiliJIDEA_Refer 🤣 NodeFactoryAnd 🤣 Solr Admin 🤣 Solr Adm	tin 🤣 swagger-core/sa 🔞 jersay j. Java, Spri 🔇 poc/MDArrayBase 📫 Yammer 🔇 Termite e2a	> using jackson 2 wl »
SciBite Search		pdeSa or heart	Drug Co-occurrences	×
 впоснемисац (19,452) вице (16,749) 	*	MEDLINE 188.885 JOURNAL 5.810 PDF 2	Things of type 'Drug' found in other sentences for your query - "pdeSa or heart". Showing 1 - 20 out of 114,314	± EXPORT
Angesternen Digette Acti	#72 448	Showing 41 - 60 out of 166,849 pdmSa or heart * About My Query	Name	Count
Doximition Visioartan Nime Onde	467 366 359	Radiosynthesis of a carbon-11 labeled PDE5 in	Cholecalciferol	19,711
Sacob/r/l/ Aldostanne	312. 325	Abstract To develop PET tracets for imaging of heart disease, a new	Ergocalciferol	18,346
fauerarenenet filter Levrengestraffine Crivation	298 247 156	Response by Blanton et al to Letter Regarding -	Glutamic Acid	16,427
Epinoprimio Hydrocortianne	124 224	Rates of Severe Early Right Heart Failure After Tille Response by Blanton et al to Letter Regarding Article, "Prei-	Dopamine Hol	13,877
Sindenselli, Cananir Cryster Gimp Christersetzbehant	186 170 171	implantation. An INTERMACS Analysis"	Ammonia (13n)	13,450
Lawcorriendan Ergonalisatemi	187 160	Myocardial Phosphodiestarases and Their Roli Abstract	Serotonin	13,115
Endettwein Shoathane Alanan	150 (47 (45	Finally, inhibition of PDE5A and PDE9A counter pathologica	Cisplatin	13,044
Show Conservations	+	Estrogen Receptor-a Non-Nuclear Signaling Co Keywards	Glutathione	12,836
T MICHORNA (1,323)	+	E2, estradio(ECa, endothelial cells,EDC, estragen dendrimer cGMP-dependent protein kinase G,PaPE, pathway-preferant	Diactic Acid Hydrocortinone	11.152
T COMPANY (1.012) T CELL LINE (682)	*	cyclic GMP:eNOS, endothelial retric conde synthase estradio	Gamma-ammobilityris Asid	10,363

Figure 4: A ranking of the drugs that co-occur most often with a target of interest

Example use cases

Extract more insights from electronic notebooks

Most Life Sciences organisations will have, at some point, deployed an Electronic Laboratory Notebook (ELN) with the goal of centralising R&D data. ELNs have become an important source of not only key experimental data points but also the development of methods and SOPs. However, much of the information stored within an ELN is captured as qualitative free text or as a collection of attachments and, as a result, the ability to mine it is typically limited to rudimentary text and keyword searches.

SciBite Search opens up new possibilities to search the ELN data more effectively and derive valuable insights. It accurately marks-up all relevant terms and concepts present within an experiment, without being limited by the indexing terms used by the ELN.

SciBite Search not only provides a better search experience for ELN data, it also enables users to answer more complex questions, such as those outlined below.

Examples of questions that can be rapidly answered with SciBite Search

- 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 for a specific target across the organisation, regardless of which synonym was used by the author of the experiment.
- Which projects are investigating potential biological therapeutics?
- 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 and what were the outcomes?
- 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?



Most ELNs only enable users to search for exactly what was written by the author of an experiment. The 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, an experiment describing work on 'muscarinic acetylcholine receptor M1' will not be found by a scientist who performs a search using the commonly used synonym 'cholinergic receptor muscarinic 1'. Even for more defined entries, the meaning of a particular field or its contents may be ambiguous, imprecise or contain multiple different data types, such as Gene, Assay Type and Species.

For more information, read the full use case.

Unlock the potential of departmental scientific documents

Much of the knowledge found within 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 search them for useful information. Even where such documents are organised in a file store, or structured in a document management system or SharePoint, such systems typically offer a very poor search experience for users, usually limited to exact matches of what was written by the document author. Similarly, inconsistent use of synonyms when writing a document makes it difficult to subsequently identify and collate all relevant data for a disease or target of interest.

SciBite Search is a fully fledged scientific search system which delivers intelligent query capabilities for individual teams or entire organisations. It's simple to deploy and easy for departments to upload their own documents whilst ensuring access is controlled appropriately.

Not only does SciBite Search make it simpler to interrogate the information managed within these 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.

SciBite Search injects semantic annotations into PDFs, whilst retaining their original format, making it easy for users to instantly get a feel for what a document is about.

For more information, read the full use case.

C SciBite Search		covid-19 X : Q DOCUMENTS SENTENCES
V FIELDS		JOURNAL ABSTRACT 75.3K JOURNAL 20.7K CLINICAL TRIAL 6.2K
DATE COMPLETED	+	Showing 1 - 10 out of 20,687 RESULTS PER PAGE: 10 V SORT: RELEVANCE V
DATE REVISED	+	covid-19 ×
E INGEST DATE	+	About My Query + Add Filter 1: Export Results Q, Advanced Search . Co-occurrences Graph
- JOURNAL TITLE	+	
📅 PUBLISH DATE	+	Coronavirus Infections
- SOURCE	+	Virus diseases caused by the CORONAVIRUS genus. Some specifics include transmissible enteritis of turkeys (ENTERITIS, TRANSMISSIBLE, OF TURKEYS); FELINE INFECTIOUS PERITONITIS; and
> AUTHORS		transmissible gastroenteritis of swine (GASTROENTERITIS, TRANSMISSIBLE, OF SWINE).
> CORRESPONDING AUTHOR		Repeated COVID-19 relapse during post-discharge surveillance with viral shedding lasting for 67 days in a recovered patient infected with
> FIRST AUTHOR		SARS-CoV-2 Relevance: 33.92 27 Jul 2020
> LAST AUTHOR		Abstract A case who revealed the longest duration of viral shedding (67 days) in current reports, presented complicated characteristic on the relapse of COVID-19 due to the inconsistent performance of chest
V ENTITIES		radiography and SARS-CoV-2-RNA detection after discharge. Lopinavir-interferon o2b boosted ribavirin following with lopinavir boosted budesonide might be a potent treatment for viral clearance.
TINDICATION (20.7K)	+	Summary A case who revealed the longest duration of viral shedding (67 days) in current reports, presented complicated characteristic on the relapse of COVID-19 due to the inconsistent performance of chest radiography and SARS-CoV-2-RNA detection after discharge.
BIOMED VERB (20.4K)	+	Discussion
= BIOMEDICAL CONCEPTS/PROCEDURES (19.9K)	+	This patient, who fulfilled discharge criteria, presented positive for <u>SARS-CoV-2</u> RNA 6 days after her first discharge, which is consistent with a previous study reported by Lan et al. However, during the second discharge surveillance, the chest CT of this patient showed new diffuse lesions in the <u>left lung</u> but <u>RTPCR</u> tests were negative (, Day 46). This phenomenon recurred on day 57 during the third hospitalization (). This finding is inconsistent with the previous study Although false-negative <u>SARS-CoV-2</u> RNA results could have occurred recarmination of viral RNA on nasopharyngeal swab and
T SPECIES (18.6K)	-	subsequent examination or main many is inconsistent with the previous attory-introduction of the subsequent examination or results of <u>RTPCR</u> testing and CT imaging (Day 50) were consistent with that of previous (Day 46), which decreased the possibility of false-results. It is difficult to find a reasonable explanation for this phenomenon. One possible explanation is that it is as a result of residual viruses in the lungs of discharged COVID-19 patients. Remaining viruses first cause pathological
Coronavirus	12.8K	reasonable explanation for this preintmention, one possible explanation is that it is as a result or resourd vitudes in the target of discharget OVID*9 patients. Remaining vitudes inst cause patient/ogleat lung changes under conflictions of weakened immunity, it will take some time for vitudes to travel to the upper respiratory tract because of circulatory disturbance which resulted from low PO2 and lung
Viruses	12.7K	dysfunction of this case. The later positive results of AMS-CoV-2 (Days 51 and 67) were further evidence, which confirm this explanation from another perspective. These findings provide new insight
Sars Virus	10.4K	grantiation of this case. The possibility that negative results of an asopharynogal swab might not synchronously reflect the presence of SARS-CoV-2 in lung issue among pointer new mong pointer
Humans	9.9K	more process of other and the process of other
Bacteria	2.9K	of this case, convalescence serological results for IgG and IgM maintained a relatively stable level since the first measurement of day 29, which suggests that serology testing could not help identify the
Orthomyxoviridae	2.4K	or the case of SARS-CoV-z in patients recovered from CoVID-19 during maintained a catarty sub-intervention and 22, which suggests that scrology recomposition reprint in the patients are strong to the second strong sub-intervention of the second strong second strong sub-intervention and second strong second strong second second strong second se
Show More		In this report, the incubation period of this case was 24 days, which is longer than 52 days reported by Li et al. The long incubation period brings great difficulties and challenges for the initial control of
Show Co-occurrences		COVID-19. The duration of viral RNA shedding after COVID-19 onset in this case lasted for 67 days, which is the longest duration of viral shedding as far as we know. Moreover, one RT-PCR test for
- ANATOMY (16.6K)	-	SARS-CoV-2 RNA was positive (, Day 51) even after five-consecutively negative results, which suggests that current criteria for discharge and termination of isolation may need to be reconsidered. Interestingly, most abnormal levels of CRP, A/G, and PO2, which indicates inflammation, liver impairment, and lung dysfunction, all turned into normal at around days 56–60 after symptom onset, and
Lung	6.3K	following with the disappearance of clinical symptoms on day 68 and sustained with the end of the observation. The exhilarating improvement of this case may benefit from the using of BUD and the
Blood	5.6K	informing with elaspleatance of unitial decay (LPV + IAI + RB), which has been reported to be ased and high effective in shortening the duration of virus shedding and alleviating symptome
		use unit of a light containation of anitring unitary (Light A light, which has been reported to be sare and right effective in anothering unitary and anitring unitary and anitring and ani

Figure 5: Summarising all Terms Identified within a document (left hand side of the screen)



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 ontologyled 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