SciBite
an ELSEVIER company

Use case

The Benefits and Applications of Democratised Ontology Management

SciBite
an ELSEVIER company

The Benefits and Applications of Democratised Ontology Management

Introduction

Ontologies play a critical role in semantic enrichment, enabling unstructured scientific text to be transformed into clean, contextualised data which can be understood and exploited by computational approaches, such as machine learning. Historically, maintaining multiple, evolving ontologies from both public and proprietary sources, has been a complex undertaking, requiring significant ontology expertise. This has presented a bottleneck for most organisations and undermines the concept that an ontology should represent a shared understanding between experts.

SciBite has developed CENtree to address these ontology management challenges and to enable ontologies to be used by a much broader audience and range of applications.

CENtree is an innovative, collaborative and centralised ontology management platform which empowers scientific subject matter experts to contribute to ontologies and revolutionises the way research-led organisations manage ontologies and serve ontologies to business and scientific applications.

Here we describe SciBite's integrated approach to simple, collaborative and robust ontology management and present example use cases of just some of the many ways which illustrate how CENtree can overcome common challenges associated with the management and application of ontologies.

CENtree: Addressing Common Ontology Management Challenges

Centralised Ontology Management

There are often several independently developed ontologies within a single domain which often contain overlapping and even conflicting information. In addition, many pharmaceutical companies also have their own proprietary lists, such as project codes and compound IDs.

CENtree uses open standards, ensuring there is no 'lock in' to proprietary language or formats. Users can upload a locally-stored ontology file and pull from public repositories whilst specifying if an entire ontology or a more specific subset of branches is imported . Users can also combine the relevant parts from one or more existing ontologies and custom lists to support the development of tailored application ontologies.

CENtree also provides version control between releases of ontologies, including the provenance of the change. For example, if users edit public ontologies, the changes can be reconciled with new releases. A merged view of the combined ontology highlights any conflicts, enabling curators to focus on things that need to be addressed.

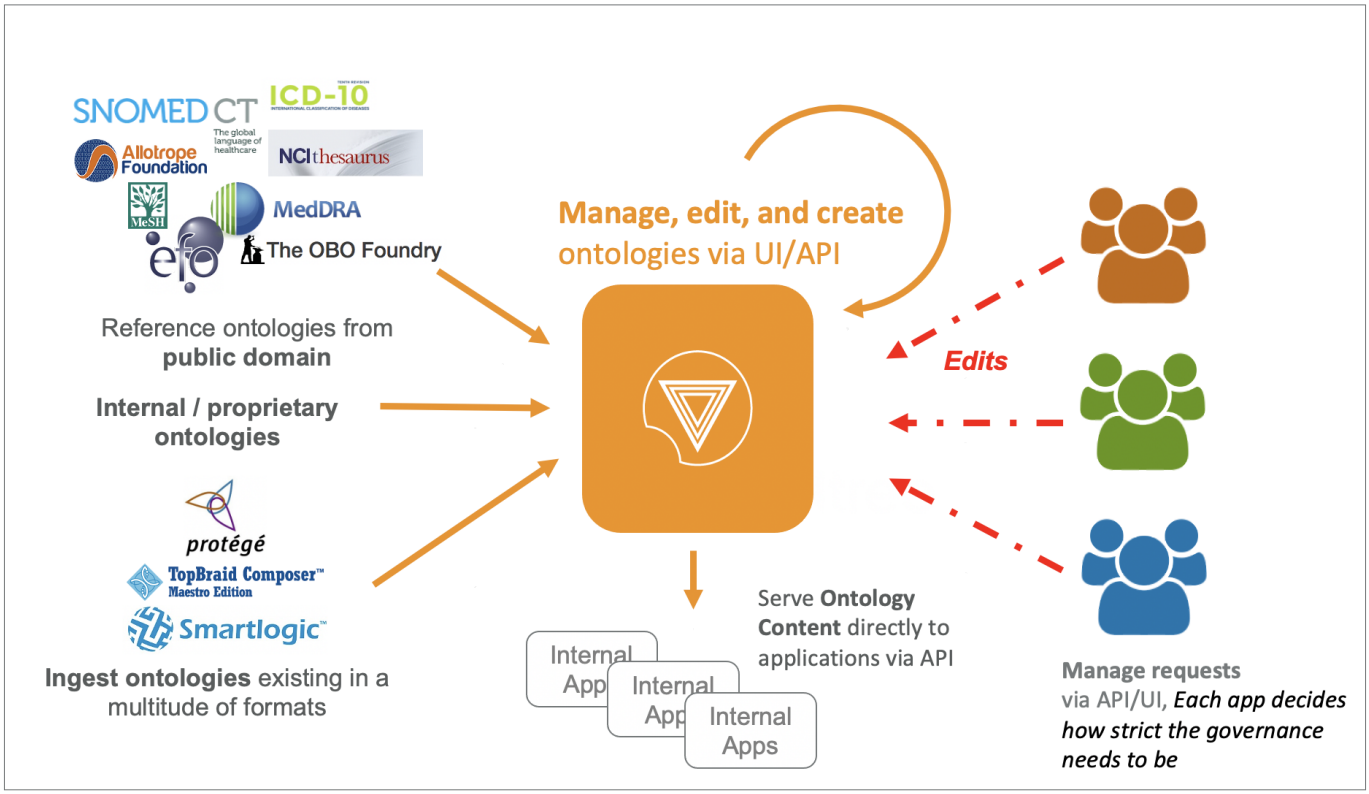

Figure 1: *Figure: CENtree: Integrated and Centralised Ontology Management*

Simplified Visualisation and Curation

As with most visualisation techniques, there is no single 'best' way to view an ontology that fits all users and all applications. Some applications benefit from a tree view, while views providing an easily digestible high-level summary, such as heat maps or bubble diagrams will make more sense for others[1].

CENtree provides users with a simple interface to navigate and make common changes to ontologies, lowering the barrier for users to contribute to editing and reducing the bottleneck of having to rely on ontology experts.

Users can add, replace or remove an ontology class, change its location in a hierarchy to reflect a new sub-class relationship, and enrich it with new metadata such as synonyms or text definitions.

1 For further examples of ontology visualisations, see: https://www.scibite.com/exploring-ontology-visualisation-techniques-for-biological-data/

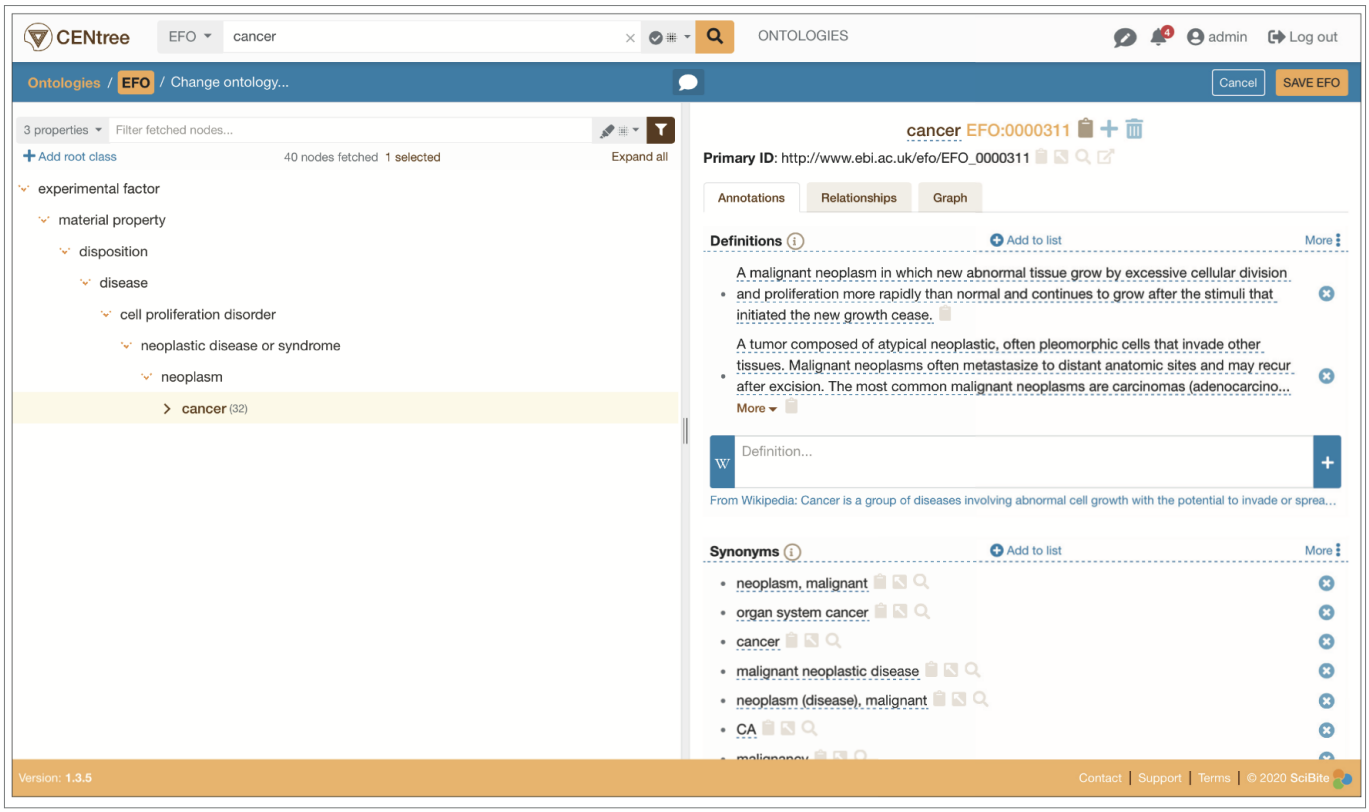

Figure 2: The user-friendly CENtree interface

Facilitating Collaboration and Integration
It is increasingly common for there to be multiple contributors, often from different locations. However, ontologies are often managed by applications that lack traceability, which hinders collaborative working. From the outset, CENtree's ontology management capabilities were built from the perspective of a multi-user environment.

For example, governance and provenance are an intrinsic part of the editing process, with the ability to manage permissions and prevent issues associated with the same term being edited by two different people simultaneously.

Ontologies can be accessed programmatically via an Application Programming Interface (API), not only by SciBite's applications, such as TERMite, our named entity recognition (NER) and extraction engine, but also by third party tools. This enables an 'ontology as a service' approach, ensuring that all applications make consistent use of up-to-date ontologies.

CENtree: Example Use Cases

Use Case 1: Next Generation Entity Data Registration Systems

The Business Problem
Pharmaceutical companies frequently use registration systems as a centralised resource of information about their Targets, In Vivo Protocols, Assays and other entities of importance to their business. These systems are often based on relational databases but, while this affords some structure to data, the associated front-end applications tend to capture data as free text fields to avoid burdening or restricting users. This lack of standardisation results in many cases where fields that are sparsely populated, and others where important information, such as the target and species, are buried in long entity descriptions. Similar entities can have totally different descriptions, making it hard to make connections and spot trends.

In addition, 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, Target and Species. Similarly, inconsistent use of synonyms during data entry makes it difficult to collate data for a disease or target of interest.

For example, Pharmaceutical companies invest heavily in technologies to perform and analyse a range of biological assays to characterise drug targets and evaluate potential therapeutic molecules. Legacy databases dedicated to managing assay data contain an amazing wealth of R&D knowledge and, as such, provide a rich resource for mining with both scientific and operational questions. However, most companies are unable to realise the true value of the data they generate.

Examples of questions that are typically difficult or impossible to answer using legacy assay data management system, such as:

- Which assays are based on fluorescence?
- Find all screens run against a specified target
- Are there any interfering compounds that show up in all assays that use a common detection technique?
- Find all protocols which use HEK cell lines

The common root cause is the way metadata is (or, more accurately, isn't) managed.

How CENtree Helps

Standardising Terminology
CENtree enables organisations to maintain an up to date ontology representing the evolving language associated with the dynamic field of modern assay technologies. Subject matter experts easily contribute to keeping things current and augment SciBite's manually curated standard ontologies (such as Assay, Cell, Gene and Allotrope), with their proprietary internal terms and vocabularies to create a single, authoritative, universally applied application ontology, for example to standardise protocol definitions in assay registration systems.

Smart Data Entry Forms
CENtree's APIs can be leveraged to make any browser-based data entry form semantically intelligent, enabling organisations to achieve semantic enrichment of their data in real-time at the point of capture. For example, a field to capture 'Tissue' can be made both semantically aware and computationally accessible without adding unnecessary burden to scientists who subsequently enter data. 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.

This ensures that assay data is consistently and unambiguously associated with key metadata.

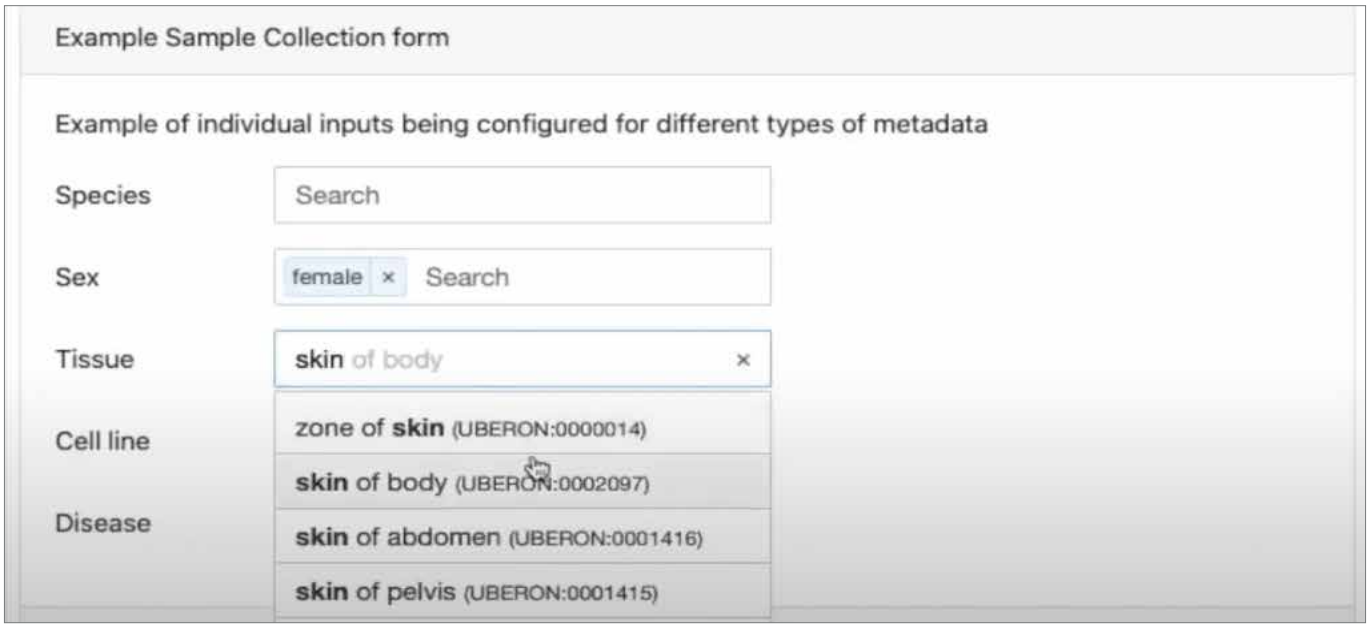

Figure 3: Semantically intelligent data entry forms

Retrospective Metadata Curation
In addition to ensuring new data is well-structured, a key consideration for pharmaceutical companies is how to align their legacy data to ontologies. Typically the process of curating and aligning such data is a time-consuming and manual process. By leveraging CENtree in combination with SciBite's scientific Named Entity Recognition (NER) microservice, TERMite, legacy data can be automatically aligned to up-to-date ontology content. TERMite can consume ontologies directly from CENtree with an easy one-click setup, enabling companies to align terminology to both public and internally-developed ontology content.

Interrogating and Integrating Data
The high degree of semantic enrichment made possible by CENtree not only makes it simpler to interrogate data, it also facilitates more complex ontology-based questions and ensures all relevant data is found, regardless of which synonym was used as the search term, enabling users to the most relevant information much faster. The resulting standardisation of data makes it far easier to integrate internal data with third party data sources, such as the Gene Expression Omnibus (GEO), ChEMBL, scientific literature, patents and commercially available databases, and ask questions across them that would have otherwise been time consuming or impossible to answer. For example, it becomes possible to compare internally generated and published data, and assess if an organisation is working on targets or diseases that its competitors aren't (and vice versa).

Use Case Summary

CENtree enables organisations to transform their entity registration systems from a data silo into a dynamic and interactive environment which provides them with a valuable source of insight. CENtree's flexible API enables it to easily integrate with any registration system and provide its semantic backbone.

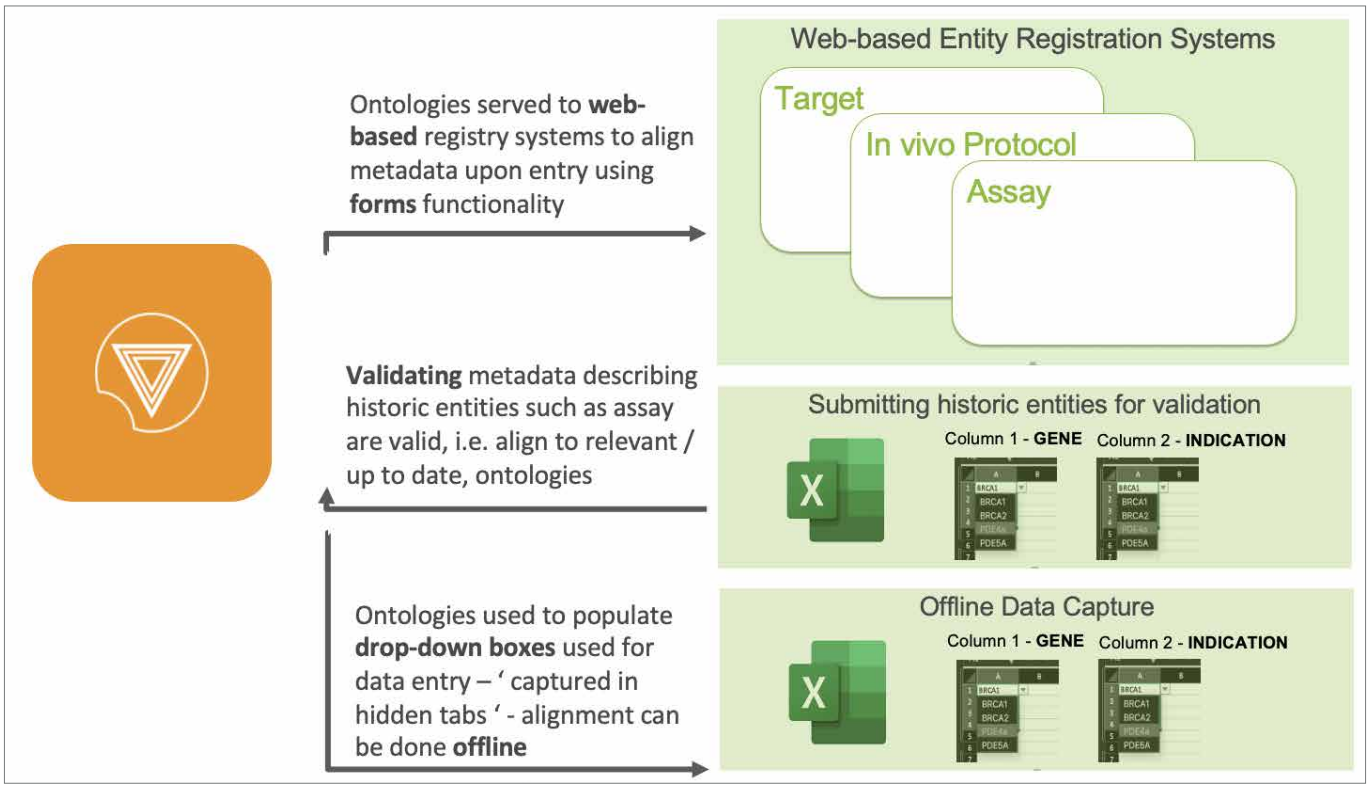

Figure 4: The Pivotal Role of CENtree in a Modern Entity Registration System

SciBite
an ELSEVIER company

Use Case 2: Facilitating the Production of Knowledge Graphs

The Business Problem
Knowledge graphs provide an increasingly important way to connect data from numerous heterogeneous internal and external data silos and provide a simple and intuitive way to represent different entities, such as genes and indications, and the relationships between them. Once the relationships between data are made in this way, it becomes easier to ask questions and make inferences that would otherwise remain unseen.

But, creating the graph is not as easy as just pulling data together. Data must first be aligned with standards, datasets need to be harmonised and relationships need to be extracted from the data and then the schema needs to be generated. Each of these aspects can present a challenge, particularly when they have historically relied on the use of expert-focussed tools, making it hard to integrate feedback from consumers of knowledge graphs back into the source ontologies and graph schema.

How CENtree Helps

Aligning with Standards
Ontologies are the foundation of any knowledge graph – they and confer real-world entities found in the scientific text with an explicit meaning and encapsulate the relationships between such entities. For example, a graph node that has the name NASH has very little use of meaning without the type of entity: is it a gene, a drug or even a person? By curating unstructured scientific text with ontologies, it can be contextualised so that it describes "things, not strings" and can be understood and used by computers.

CENtree provides a centralised resource for ontology management and enables users to extend ontologies, such as SciBite's extensive VOCabs, manage internal vocabularies, such as compound IDs and study codes, or develop new ontologies for domains not currently captured in a VOCab.

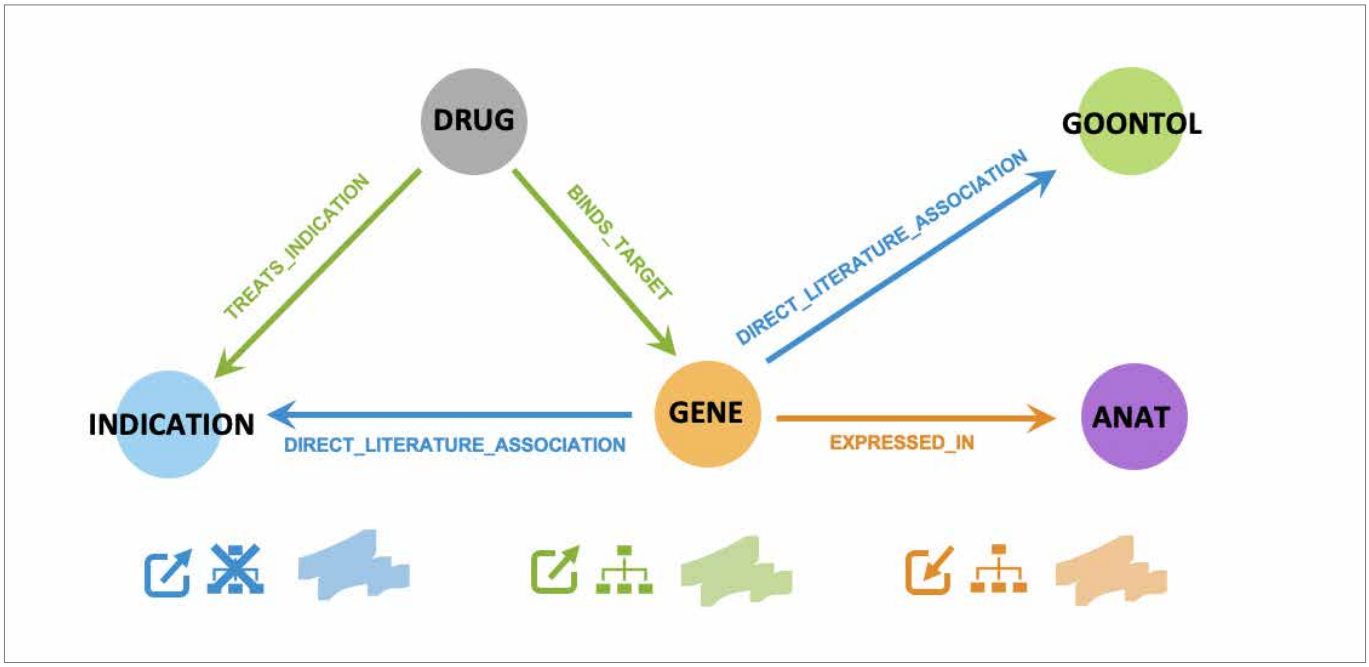

Figure 5: A visual representation of the relationships between a selection of scientific entities

SciBite
an ELSEVIER company

Harmonising Datasets
The ability to create knowledge graphs is critically dependent on the ability to integrate data from multiple sources. The combination of CENtree and SciBite's Named Entity Recognition (NER) engine, TERMite, enables the rapid identification of scientific entities within unstructured text, regardless of the synonym used by the author. TERMite aligns these entities to single unique identifiers captured in our ontologies, resulting in 'clean', structured data that can be integrated with other sources.

But ontologies deliver much more than data harmonisation. One of the roles of an ontology is to provide a common model of knowledge associated with a given domain so, for example, the fact that Type II Diabetes Mellitus is an endocrine disease is already encapsulated within the ontology that is used to enrich the source text.

Once a disease entity has been harmonised to a single ID, such as the MeSH ID, it makes it possible to map it to other representations of the disease from other ontologies such as EFO (Experimental Factor Ontology), OMIM (Online Mendelian Inheritance in Man), or SNOMED (Systematized Nomenclature of Medicine). This enables information found in the literature to be augmented with additional information from other structured data sources. For example to find drugs used to treat that indication from ChEMBL or to identify genes associated with the indication of interest from OpenTargets. Essentially these linkages provide a 'springboard' for further exploration across the knowledge graph.

Extracting Relations from Data
A common challenge is to identify when a specific association exists between two entities rather than when they are simply just being mentioned in the same document. TExpress to define semantic patterns, or groups of patterns, which describe a relationship between two concepts, such as a gene and drug, in the form Gene-Verb-Drug. These can can then be extracted from the text as semantic triples, aligned to ontologies managed by CENtree. Ultimately this process can generate a set of the various attributes that describe a relationship or association which can be ingested into, and subsequently enrich, a knowledge graph.

Schema Generation
Schema Generation is the creation of a high level meta graph of the relevant entities and the relationships between them. CENtree can be used to create a simple representation using an initial 'bridging ontology' which can then be enriched with more ontologies, such as a disease entity populated by EFO disease classification.

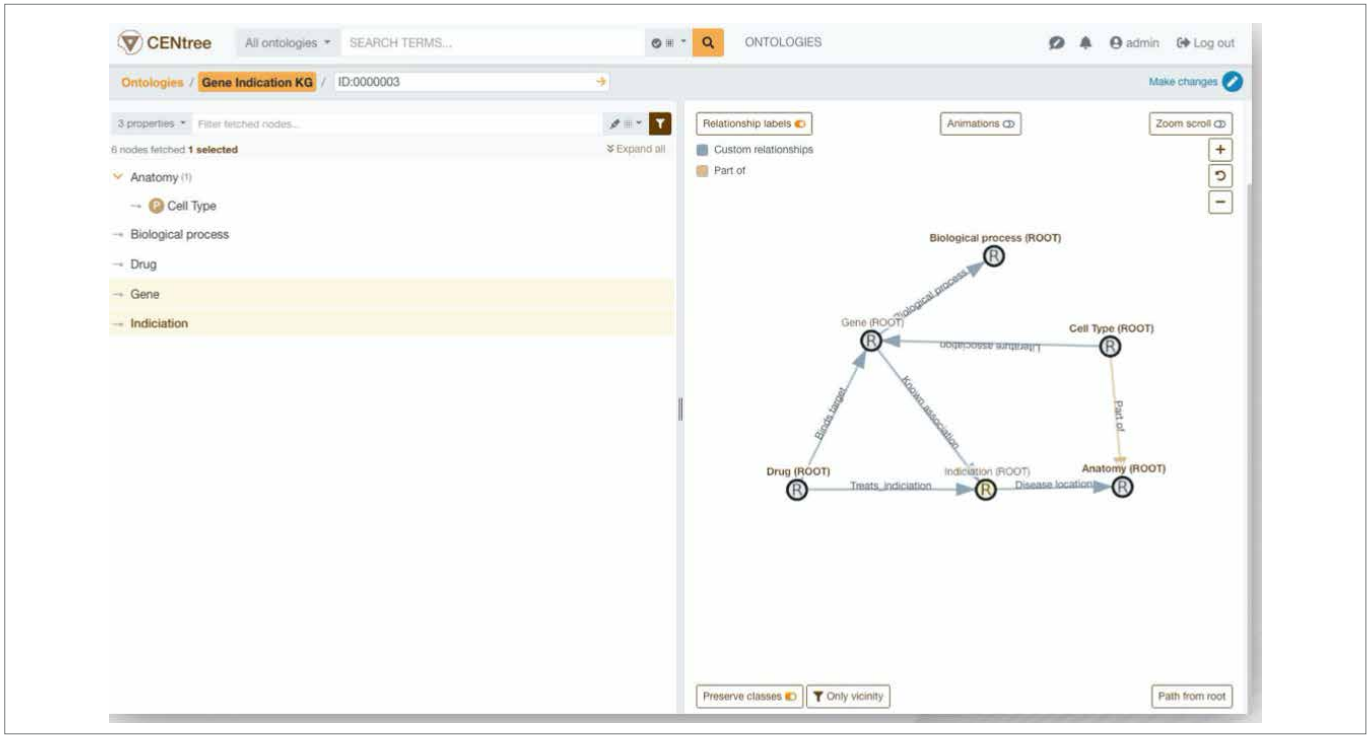

Figure 6: A meta graph representation in CENtree

Once the schema has been generated, CENtree enables you to export it to your graph database of choice in whatever format suits your particular application. For example, if you are generating an enterprise graph to hold large, normalised datasets from across your organisation that can be retrieved by other systems, then the schema can be exported to an RDF Triplestore. Whereas if your graph is designed to support investigative analytics as part of a target validation or drug repositioning initiative, then you may prefer to export your schema in JSON format so it can be ingested into a more intuitive labelled property graph.

Use Case Summary

CENtree enables organisations to broaden their community of practice when it comes to creating and updating knowledge graphs, providing a portal to enable contributions from subject matter experts as well as experienced ontologists. Since SciBite's technologies don't need to be used in isolation – they are all provided as easy-to-consume microservices which can be easily embedded into an automated knowledge graph creation pipeline.

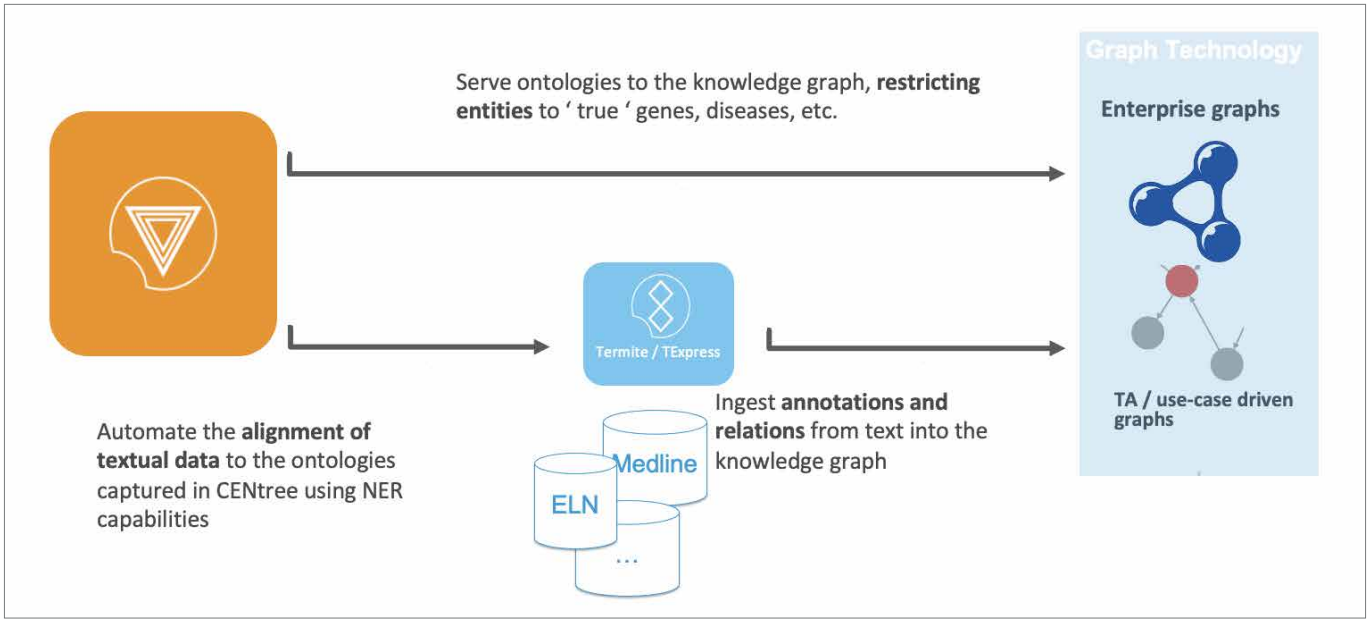

Figure 7: *The Pivotal Role of CENtree in a Knowledge Graph Creation*

SciBite
an ELSEVIER company

Additional Use Cases

CENtree is by no means limited to the use cases described above. As illustrated below, CENtree can play a pivotal role in a range of use cases, including:

ELN and LIMS Enrichment
Embedding ontologies directly into ELN/LIMS data entry forms to enable prospective alignment of experiment metadata to ontologies and aligning historical ELN/LIMS experiments to relevant standards.

Text/Data Analytics
Automating the alignment of textual data to the ontologies captured in CENtree using TERMite's Named Entity Recognition (NER) capabilities for applications such as:

- Pharmacovigilance
- Drug Repositioning
- Horizon Scanning
- Phenotype Triangulation

Artificial Intelligence/Machine Learning (AI/ML)
Using ontologies managed within CENtree to generate high quality training data sets, as well as using AI/ML to build de novo candidate ontologies within CENtree.

Harmonisation of Disparate Sources
Aligning the textual data to common terminology to ensure data is Findable, Accessible, Interoperable and Reusable (FAIR), regardless of its source

Semantic Search
Generation of a semantic index to bring scientific understanding to enterprise search

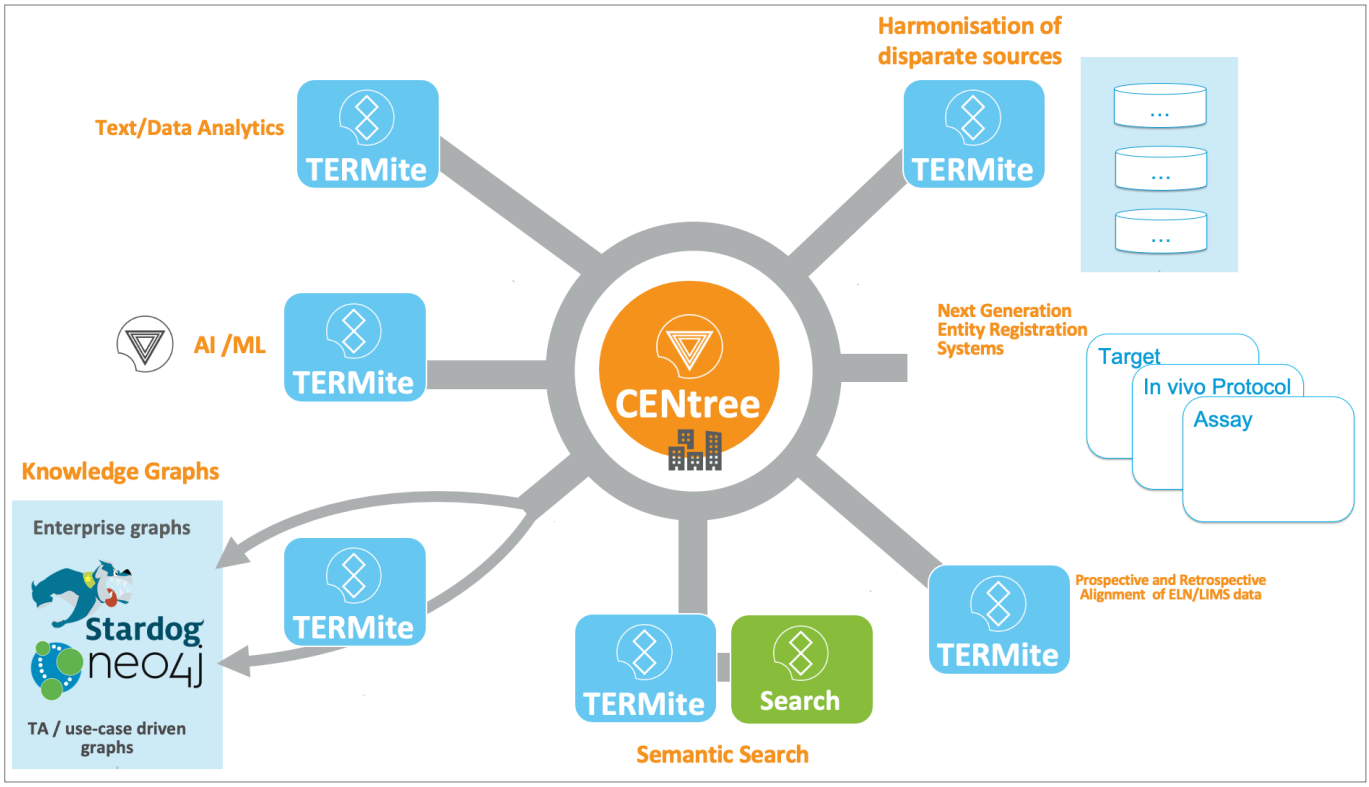

Figure 8: A selection of just some of the many use cases for CENtree

SciBite
an ELSEVIER company

Summary

The challenges associated with managing an evolving ontology are amplified when organisations rely on multiple ontologies from different sources. CENtree addresses this problem with simple to use ontology management capabilities built on a robust enterprise-ready platform.

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.

SciBite
an ELSEVIER company

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