

Use case

**More than FAIR:**  
Unlocking the value  
of your bioassay data







# More than FAIR: Unlocking the value of your bioassay data

## Introduction

One of the most valuable assets for any organisation is its data. However, most pharmaceutical companies are unable to realise its true value as a result of either i) deploying a data management system that is geared towards entering rather than mining data and/or ii) replacing such systems over time, resulting in silos of legacy data.

### Common Issues Limiting the Use of Data

- Data management system has basic search tools with rudimentary capabilities
- Data fields containing unstructured text or ambiguous information
- Data has been collected using different systems which lack consistency (e.g. different formats and/or capture different metadata)
- Reliance on a commercial system that is no longer supported
- Legacy data cannot be easily migrated to a new or upgraded database
- Lack of API to query the data using dedicated search tools

The way in which an organisation captures and manages its data is fundamental to addressing this problem. A wider scientific community initiative has resulted in the establishment of the FAIR principles<sup>1</sup> to ensure that data is Findable, Accessible, Interoperable and Reusable. Although initially focused on the accessibility of public domain data, the FAIR principles are rapidly gaining interest from the pharmaceutical industry<sup>2</sup>.

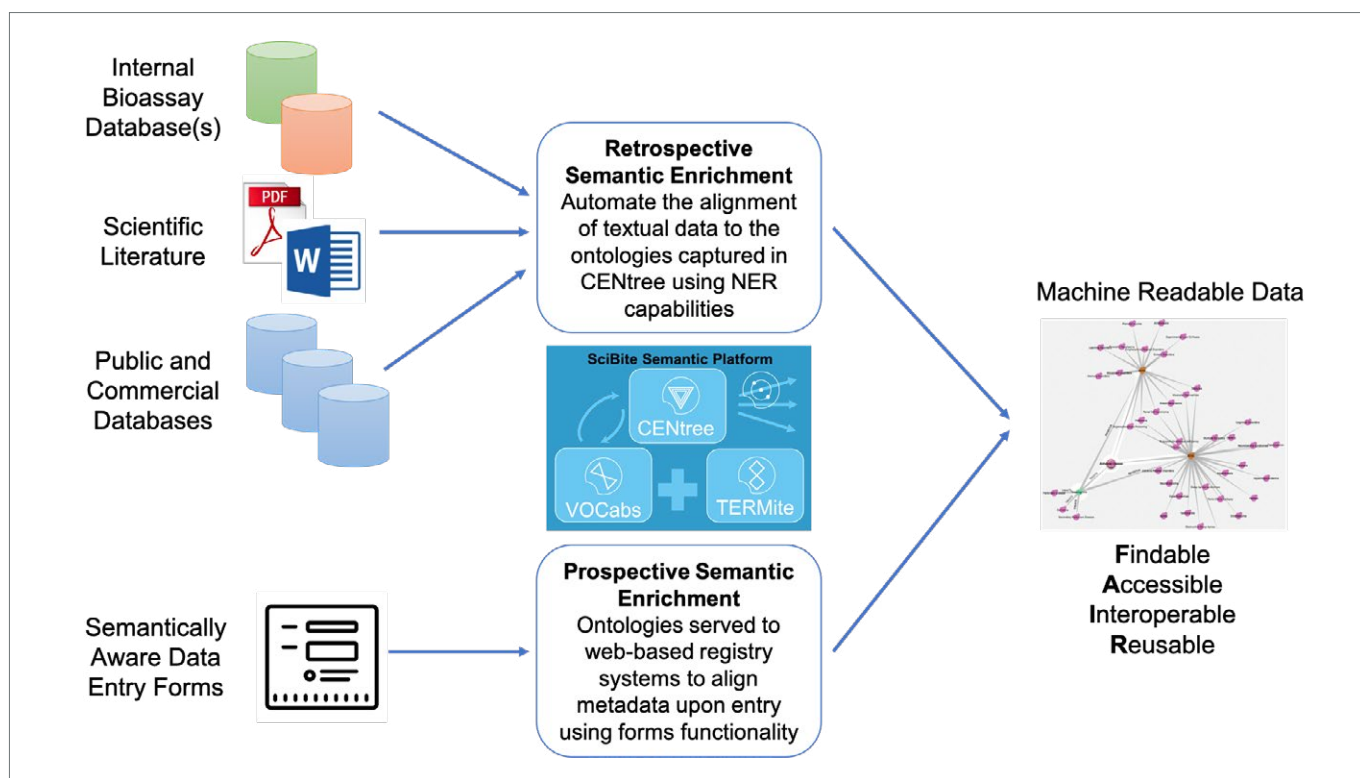
The benefits of FAIR can be illustrated using the example of bioassay data management. A significant proportion of the pre-clinical data that has been accumulated by every pharmaceutical company is a result of conducting a range of biological assays to characterise drug targets and evaluate potential therapeutic molecules. Databases dedicated to managing bioassay data contain an amazing wealth of R&D knowledge and, as such, provide a rich resource for mining with both scientific and operational questions.

### SciBite: semantically enriching bioassay data

When implementing a change in data management strategy, it should not be limited to legacy data. Based on FAIR principles, SciBite uses semantic enrichment to unlock the value of bioassay data via retrospective analysis of existing data and via semantically intelligent forms to enrich new data at the point of capture and ensure that it is 'born semantic'. This ensures that all of an organisation's bioassay data can be queried easily and intelligently (Findable), is represented such that it is machine-readable (Accessible), can be integrated with other data sources (Interoperable) and is open, richly described and understandable (Reusable).

<sup>1</sup> Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18

<sup>2</sup> As exemplified by the formation of a new dedicated track at Bio-IT World 2018: 'FAIR Data for Genomic Applications'. See <http://www.bio-itworldexpo.com/genomic-fair-data/>



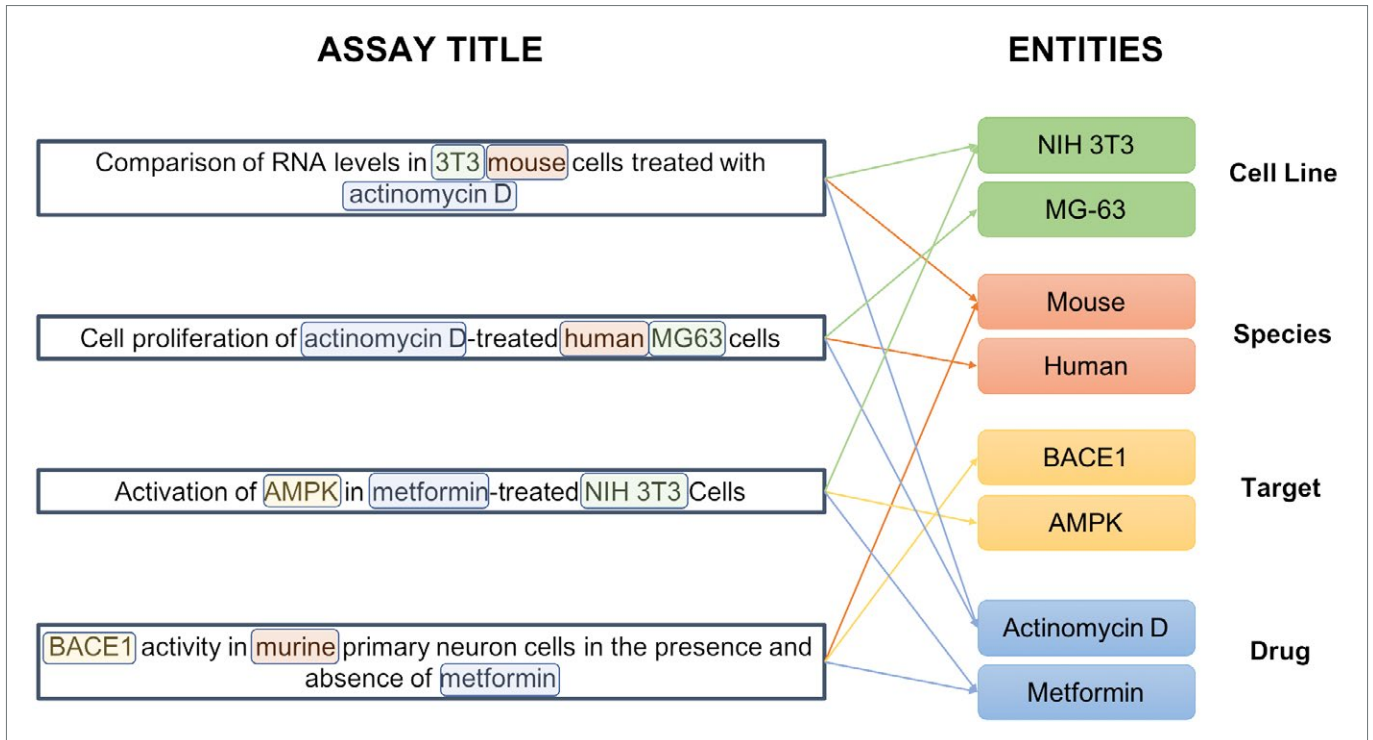
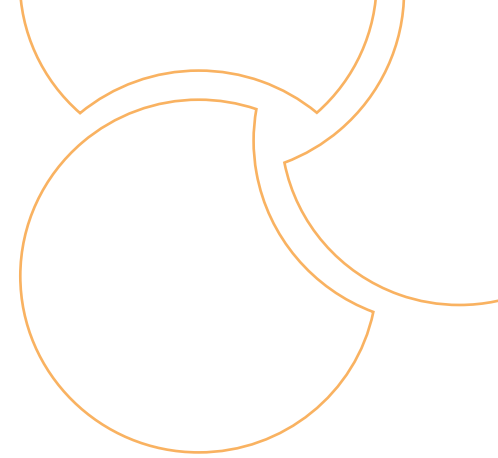
**Figure 1:** The Role of the SciBite Platform in enabling the retrospective and prospective semantic enrichment of a range of disparate, unstructured data sources

Bioassay data management systems are often based on relational databases. 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. 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, a search of a typical bioassay database for the Alzheimer's related gene, PSEN1, would miss references to synonyms such as Presenilin-1, AD3 and PSNL1.

To ensure data is interoperable, the source of ontologies is a key consideration - as the use of a proprietary ontology results in reliance on a specific vendor. By using public standards, such as BAO (BioAssay Ontology), ChEMBL (chemical entities), CLO (Cell Ontology), EFO (Experimental Factor Ontology), GO (Gene Ontology), AFO (Allotrope Foundation Ontology), MeSH (Medical Subject Headings), NCBI Species and human gene identifiers from the HGNC, the resulting enriched data is open and transportable from system to system, which is fundamental to adhering to FAIR.

SciBite's ontology management platform, 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, or VOCabs, (such as Assay, Cell, Gene and Allotrope), with their proprietary internal terms vocabularies to create a single, authoritative, universally applied terminology to standardise the way assay protocols are defined. CENTree also leverages machine learning techniques to suggest potential new terms that are similar to other assay-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 bioassay data, resulting in a semantic index endowing text entries an explicit, specific meaning.



**Figure 2:** Extraction of Cell Line, Drug, Species and Target entities within the unstructured titles of a selection of assays. The resulting semantic index enables connections to be made between bioassays

In addition to unlocking the potential of legacy data through retrospective analysis, 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 'Species' 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.



**Example Sample Collection form**

Example of individual inputs being configured for different types of metadata

Species	<input type="text" value="Search"/>
Sex	<span style="border: 1px solid #ccc; padding: 2px;">female</span> <span style="border: 1px solid #ccc; padding: 2px;">×</span> <input type="text" value="Search"/>
Tissue	<span style="border: 1px solid #ccc; padding: 2px;">skin of body</span> <span style="border: 1px solid #ccc; padding: 2px;">×</span>
Cell line	zone of skin (UBERON:0000014)
Disease	skin of body (UBERON:0002097)
	skin of abdomen (UBERON:0001416)
	skin of pelvis (UBERON:0001415)

**Figure 3:** *Semantically Intelligent Data Entry Forms*

### The benefits of semantically enriched data

By simplifying data integration, SciBite makes it possible to gain a holistic view of everything that is known, both internally and externally, about any compound, target or disease of interest. For example, given one or more chemical compounds in SMILES format, retrieve the associated activity data including compound structure, activity type, activity value, activity units, assay type, assay description, target information including the gene symbol and PubMed ID.

Semantic enrichment not only makes it simpler to interrogate bioassay 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.

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

- 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 assays have utilised a rodent cell line?
- Which protein kinases have we run screens for (and how many screens have we done for each one)?
- Which experimental techniques are growing across the organisation and would benefit from a core facility?

In addition to bioassay data, SciBite can also apply semantic enrichment to both internal and external gene expression data (such as the Gene Expression Omnibus, GEO) and other third party data sources such as scientific literature, patents and commercially available databases<sup>3</sup>. Neither is this limited to pre-clinical data – it is also possible to enrich and unlock the value of clinical data stored in SAS files.

<sup>3</sup> See SciBite's library for examples of semantic enrichment of the biomedical literature

Since the resulting data will be as well structured and interoperable as public data such as ChEMBL and PubChem, it becomes facile to integrate multiple disparate sources 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 assay data, and assess if an organisation is working on targets or diseases that its competitors aren't (and vice versa).

### **Summary**

One of the key challenges faced by the pharmaceutical industry is making scientific data Findable, Accessible, Interoperable and Reusable.

SciBite's unique combination of retrospective and prospective semantic enrichment immediately brings

scientific intelligent search to any bioassay platform, enabling the data within it to be readily integrated with other sources and making it computationally accessible for automated analysis.

SciBite enables the wealth of information locked in bioassay databases to be exploited effectively and efficiently.

### **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.









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](mailto:contact@scibite.com)



**Head Office:**

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

 [www.scibite.com](http://www.scibite.com)  
 [contact@scibite.com](mailto:contact@scibite.com)  
 LinkedIn: SciBite  
 Twitter: @SciBite  
 +44 (0)1223 786 129