SciBite

Use case

More than FAIR: Unlocking the value of your bioassay data

SciBite

More than FAIR:
Unlocking the value of your bioassay data

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 (see Figure 1).

- 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

Figure 1: *Common Issues Limiting the Use of Data*

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[1] 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[2].

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 SciBite Forms: an intelligent data entry solution for newly generated data. This ensures that all of an organisation's bioassay data can be queried easily and intelligently (Findable), is represented such that it can be used computationally (Accessible), can be integrated with other data sources (Interoperable) and is open, richly described and understandable (Reusable).

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.

1 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

2 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/

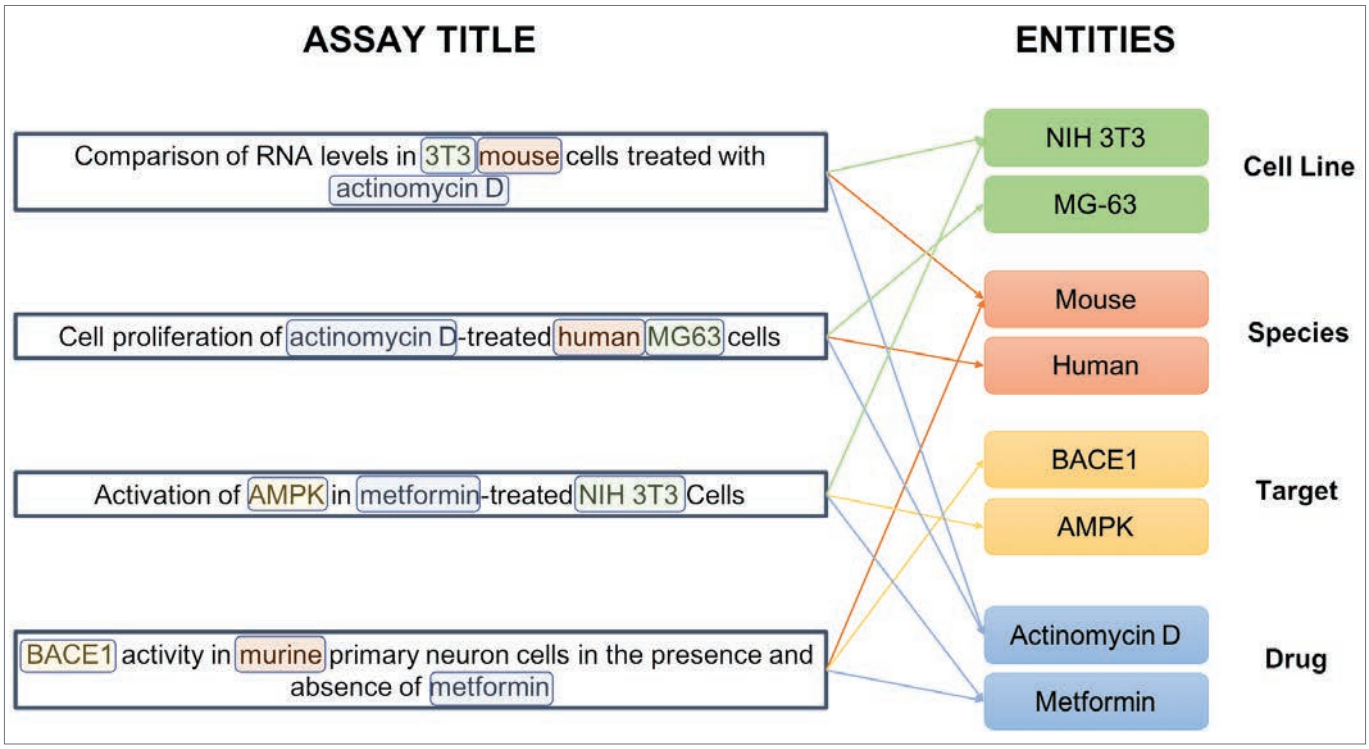

Figure 2: *Extraction of Cell Line, Drug, Species and Target entities within the unstructured titles of a selection of assays. The resulting semantic index enables connections to be made between bioassays*

SciBite applies standard, well established ontologies and controlled vocabularies to bioassay data, resulting in a semantic index endowing text entries an explicit, specific meaning. 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.

In addition to unlocking the potential of legacy data through retrospective analysis, SciBite can also make any browser-based data entry form semantically intelligent through SciBite Forms, enabling organisations to achieve semantic enrichment of their data in real-time at the point of capture. For example, by leveraging this capability, the 'Species' field 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.

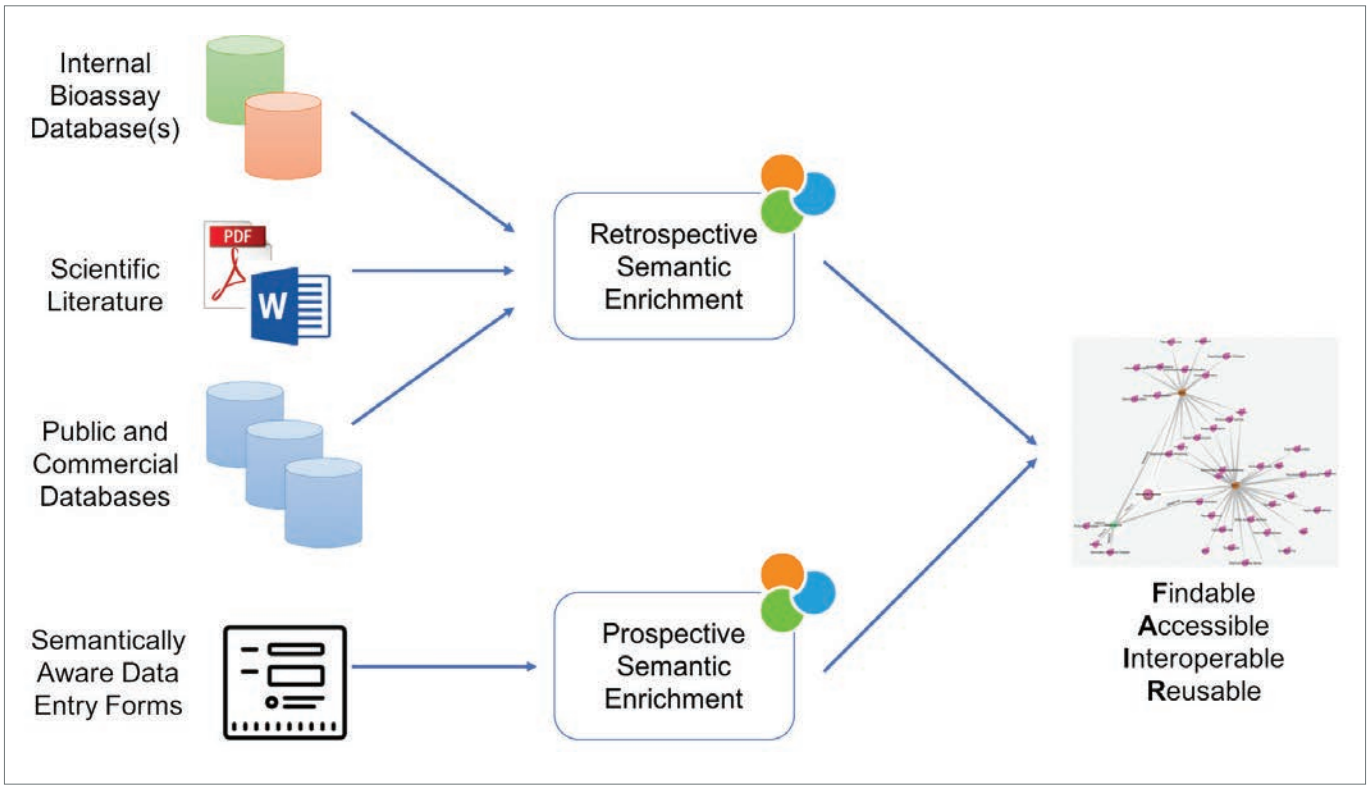

Figure 3: Retrospective and prospective semantic enrichment of a range of disparate, unstructured data sources

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 (see Figure 4) and ensures all relevant data is found, regardless of which synonym was used as the search term.

- 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?

Figure 4: Examples of questions that can be rapidly answered with semantically enriched bioassay data

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[3]. 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.

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's award winning[4] platform is the culmination of tens of years of experience applying Semantic Analytics to pharmaceutical data. Global pharmaceutical companies and emerging biotechs have partnered with SciBite and are leveraging our unparalleled know-how to unlock the potential of the wealth of unstructured scientific data at their disposal.

3 See SciBite's library for examples of semantic enrichment of the biomedical literature

4 SciBite has been recognised with a series of awards, including Bio-IT World's Best of Show 2017 and the British Chamber of Commerce in Japan's 2017 British Business Award for Innovation. The latter is in recognition of our transformation of data management in the life sciences, and the opportunity this has brought for Japan to gain a global advantage in the sphere.

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

THE LANGUAGE OF SCIENCE

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