SciBite
an ELSEVIER company

Use case

Eliminating the data preparation burden

SciBite
an ELSEVIER company

Eliminating the data preparation burden

The era of data-driven R&D is motivating investment in technologies such as machine learning and natural language processing to provide deeper insights into new drug development strategies. Despite major advances in technology, many computational approaches struggle to deal with the complexity and variability of unstructured scientific language. One fundamental of data science remains unchanged: the accuracy and reliability of results are both critically dependent on clean, high quality data.

However, the data cleansing and annotation work required to achieve clean, high quality data can be costly, often prohibitively so. For example, data scientists spend almost 80% of their time as 'data janitors', collecting, cleaning, formatting and linking data, and only 20% of their time actually analysing data. Furthermore, for most data scientists, data preparation is the least enjoyable part of their role. This presents a significant risk: when people spend a significant part of their time on a task they don't enjoy, mistakes are bound to occur.

For most Pharmaceutical companies, extracting insight from heterogeneous and ambiguous data remains a challenge, consuming a significant amount of the time of their already constrained data scientist resources.

Common issues with semi-structured data

Much of the data within a Pharmaceutical company is managed within databases and spreadsheets. However, the fact that such systems provide a degree of structure shouldn't be confused with meaning that the data is clean.

For example, different systems can use different field headings to represent the same biological concept and/or be constrained using different vocabularies, resulting in different representations of the same thing. Such inconsistencies between data sources is a barrier to data integration.

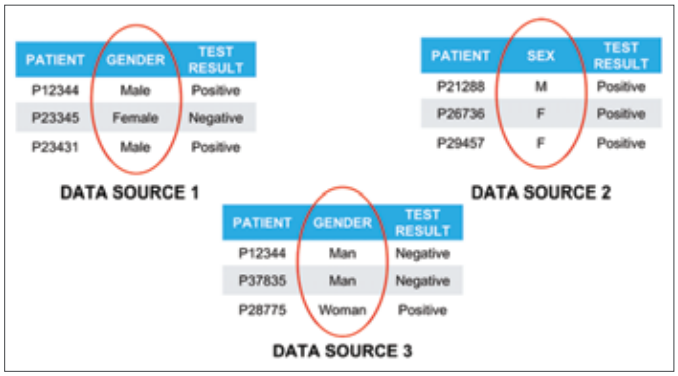

Figure 1: *An example of 3 different data sources, each with a different approach to capturing gender information*

Similarly, database fields that enable entry of free-text information typically result in different users entering information in different ways or to different levels of detail. For example, a free-text field 'Target' might be populated with the name, ID or description of the target, internal abbreviations that are not transparent, placeholder values (or reminders such as "ask Senior Scientist to fill this in"), some combination of these or even nothing at all. While such entries may be understood by an individual, they create a barrier to automating the aggregation and linkage of data, restricting the future use of the data and therefore limiting its value.

ASSAY ID	ASSAY METHOD	TARGET
A-78899	Fluorescence	P15056
A-79990	Cell Proliferation	BRAF
A-89101	Fluorescence	B-Raf proto-oncogene, serine/threonine kinase
A-92733	Dose Response	*Enter Target Here*
A-93062	Fluorescence	The target is PTEN (phosphatase and tensin homolog) which is being studied because…

Figure 2: *An example of inconsistent use of a free-text data field*

Even where some degree of consistency exists, multiple terms can be used to describe the same thing, such as 'heart attack', 'myocardial infarction' and 'serious heart event', making it difficult to find all references to a term of interest.

These are just a few examples – there are many different ways that 'messy' data is manifested within a typical Pharmaceutical company. Ultimately, its presence hampers downstream data integration, searching and analysis.

SciBite: automated data cleaning

SciBite's award winning 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 scientific data at their disposal.

Recognising the need to simplify the process of cleaning data for our customers, SciBite has acquired FactBio and integrated its Kusp technology into the SciBite platform. By combining the unrivalled depth and breadth of SciBite's ontologies with Kusp's user friendly data curation capabilities, SciBite provides the tools to automate data cleansing and generate the high-quality data needed for meaningful analysis.

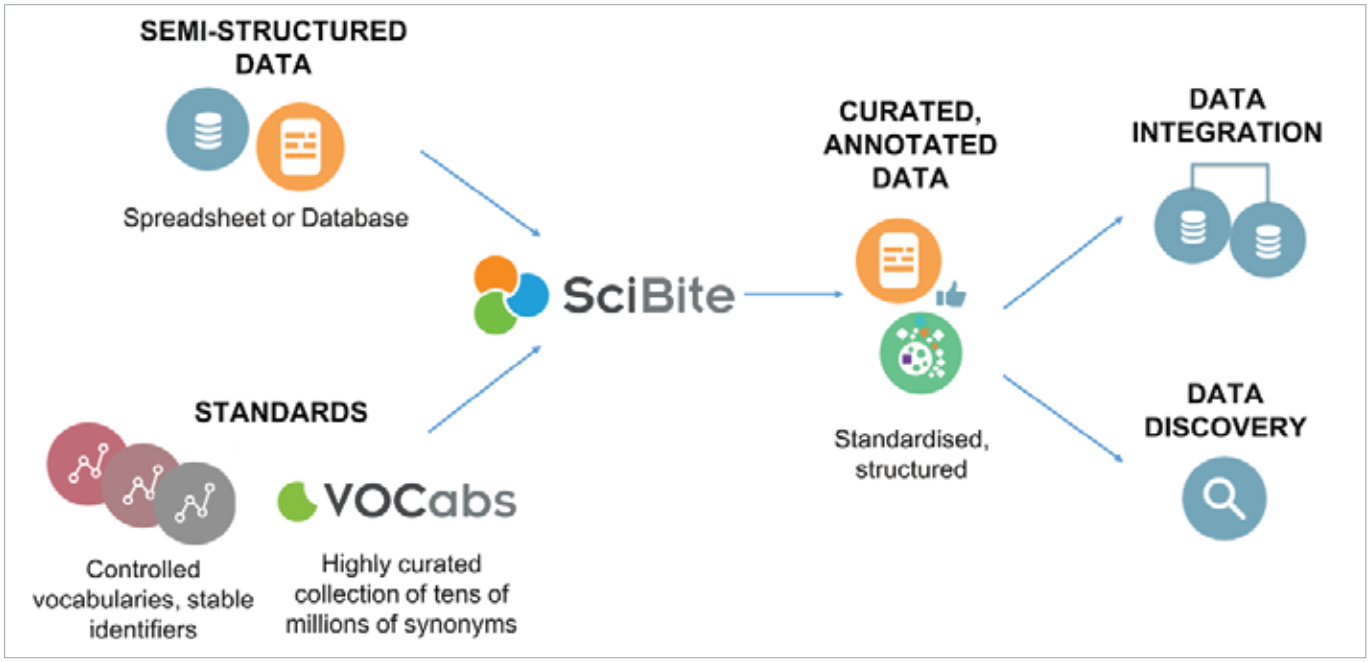

Figure 3: *Through integration of FactBio's Kusp, the SciBite platform automates cleaning and standardisation of semi-structured data, facilitating downstream data integration and discovery*

Accurate, automated data annotation

Initial data standardisation is achieved via a proprietary algorithm, which is designed to exploit ontologies to automate the semantic enrichment and annotation of data, whilst also coping with ambiguities such as synonyms or typographic errors. This is not limited to public reference ontologies – users can also leverage SciBite's manually curated ontologies or a combination of both.

Since many data sources, including spreadsheets, rarely contain a single type of data, SciBite provides the flexibility for users to specify which ontology (or combination of ontologies) to apply to a specific column. It is easy for users to assess which ontology, or combination of ontologies, provides the best coverage for your data, taking into account the desired level of accuracy.

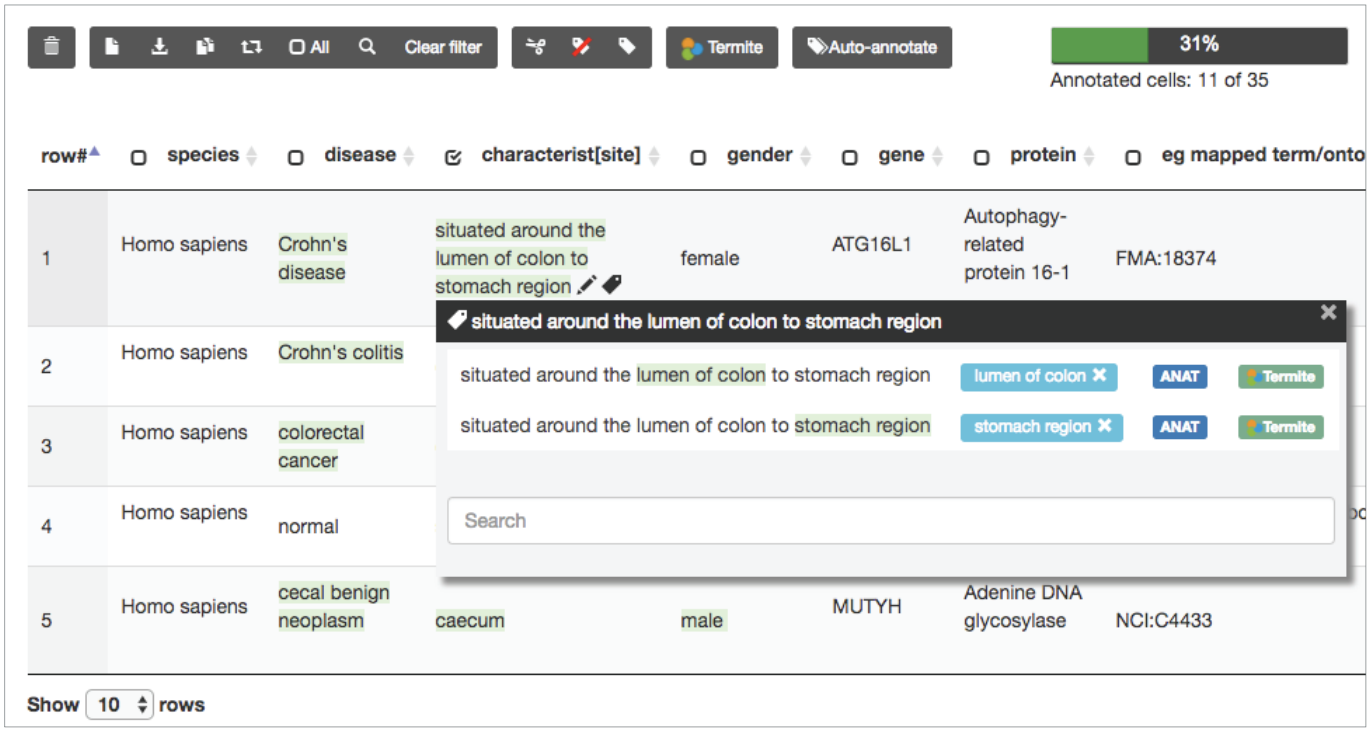

Figure 4: Simplified manual annotation via a user-friendly interface

SciBite ensures a balance between throughput and accuracy. Following the initial automated process, annotations can be updated both individually and in bulk through an intuitive interface which displays ontologies alongside the data to be annotated, eliminating the need to switch between applications/websites.

Dealing with 'messy' data

Proprietary information, such as project codes and cell line IDs developed for internal use, can confound most annotation algorithms. SciBite provides a simple form to enable users to create new business rules to guide the annotation process and improve the end result. It is not always possible to find identical matches within ontologies, so users can set confidence levels for automatic processing with different degrees of fuzziness.

For particularly messy data sets, 100% annotation coverage may not be realistic. Users can set annotation targets, which can be reached by a combination of manual and automated curation, as well as setting desired confidence levels. Users can easily identify which data is unannotated and the number of occurrences of different entity types.

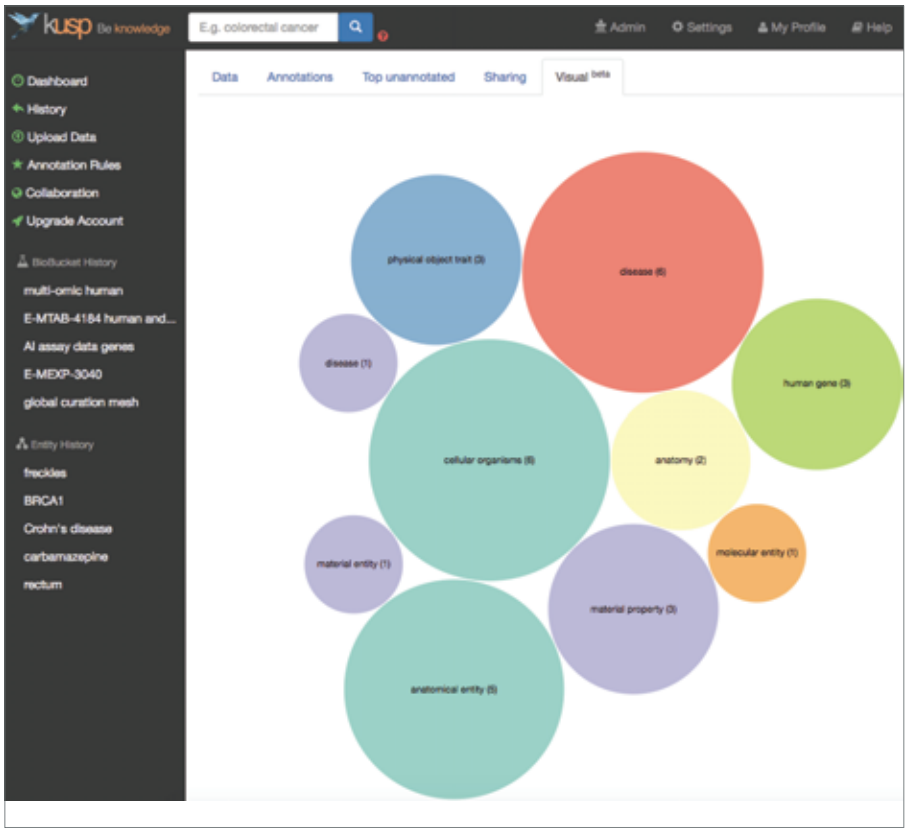

Figure 4: A visual illustration of the relative occurrence of different entity types in the dataset

Data sharing made simple

To support collaborative curation projects, standardised data can be easily disseminated to colleagues while ensuring that data owners have full control over who can see, edit or annotate it. Key properties are maintained along with each annotation, including details of which ontology was used to create it and the specific version of that ontology that was used. Annotated data can be exported in a number of formats including JSON, Excel and TXT, facilitating integration with third party systems.

All of the core functions can be accessed programmatically, enabling the upload, annotation and download of data to be automated. The results of automated processes are made transparent via a web interface, ensuring users are in full control throughout the curation process – first, by defining the level of annotation "confidence" (i.e. how much "fuzziness" is allowed) and second, by spot-checking annotation results via the visual at-a-glance summary.

The flexible combination of a powerful REST application programmatic interface (API) and an easy to use interface, provides users with control over both curation quality and throughput.

Example use case – Retrospective curation of heterogeneous metadata

The Business Challenge:
A global Pharmaceutical company needed to collate a large volume of metadata from internal spreadsheets and convert the annotated data to JSON format for ingest into their data integration platform.

The data sources were heterogeneous, lacking consistency both in terms of the headings used for spreadsheet columns and their contents, particularly for free text entries. While some of the metadata, such as species names, were unambiguous and well-structured, the naming of cell lines, diseases and gene was less clear-cut. For example, the same disease was recorded using different synonyms, cell lines were often referred to using internal codes and genes were recorded as either gene symbols or identifiers. In addition, many entries also contained typographic errors, such as "colorectal cancerc" for "colorectal cancer".

Given the volume of data, manual curation would be very labour-intensive and time-consuming.

The SciBite Solution:
To facilitate data integration, all metadata was curated and aligned to a common standard. Over 10,000 gene names/ identifiers in multiple spreadsheet columns were annotated to the Ensembl stable identifiers. Disease names were annotated with unique ontology terms from the Disease Ontology and, while species information was unambiguous in the dataset, it was mapped to the NCBI Taxonomy to ensure that it was computationally accessible by the data integration system.

Business rules were formulated to enable internal cell line codes to be aligned with ontology terms from the Cell Line Ontology and annotate over 1,000 cell line names with high precision.

The entire curation process was achieved in a matter of minutes, delivering considerable time savings as well as significant improvements in data quality.

Business Benefits:
• Curate thousands of metadata values with high throughput and precision
• Avoid copy-and-pasting between data source spreadsheets and Kusp, eliminating the
chance of errors or data corruption
• Export annotated data in JSON format, enabling transfer to 3rd party data integration
systems

Example use case – Import and annotation of public data

The Business Challenge:
A large biotech company wanted to curate a group of human tissue and cell line gene expression experiments related to two diseases of interest and align the annotation to globally referenced Identifiers in standard ontologies, specifically ChEMBL for drug terms, MeSH for diseases and CLO and EFO for cell lines. Since no single database contained all the required data sets, the data had to be sourced from three different sources (NCBI Gene Expression Omnibus (GEO), EMBL-EBI ArrayExpress and EMBL-EBI Expression Atlas).

The data from these sources had already been curated, but to different extents or using different standards. In some cases, many text-based data fields were not constrained, resulting in the fields being used inconsistently between experiments. For example, GEO lacks a field for Disease, so for some experiments disease information is found in the Description field, while in others it is found within the Sample Name field.

The challenge of manually curating this data was compounded by the size of the datasets. For example, GTEx contained over 100,000 metadata values in approximately 19,000 spreadsheet rows.

The SciBite Solution:
Source metadata from the three databases were downloaded via their respective FTP or REST API. The Kusp API was able to automatically load the datasets and run a first-pass, high precision annotation within a few minutes before presenting the results in the web interface for checking. Where errors were identified, such as the cell line DOV13 being misspelled "DOR13" in the Genentech cancer cell line data, bulk corrections were easily made via the user interface. In some cases, business rules were formulated to increase annotation accuracy.

To illustrate the results, in the original CCLE dataset from Expression Atlas, 57% (537) of the 934 cell lines names had been manually annotated to EFO ontology terms. Using a combination of EFO and CLO, Kusp achieved a significant improvement the annotation coverage to 89% (832) without the need for manual curation. Kusp was run with a high confidence level, giving assurance that the annotations for the extra 283 cell line terms were accurate.

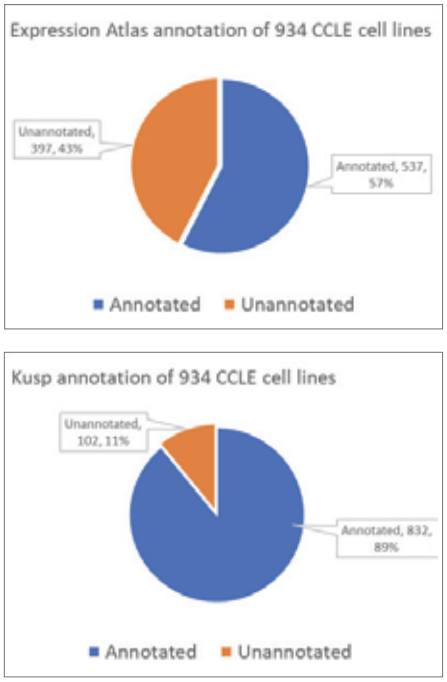

Figure 6:
A comparison of the annotation coverage for Cancer Cell Line Encyclopaedia (CCLE) cell lines before and after annotation by Kusp

SciBite
an ELSEVIER company

Business Benefits:

- Rapidly harmonise annotations to standard ontologies
- Annotate large datasets, such as human gene expression experiments, with high precision in a matter of minutes
- Improve annotation coverage and accuracy with user-definable business rules
- Scalable to tens of thousands of experiments

Summary

With the acquisition of FactBio, SciBite delivers an integrated, cost-effective solution to significantly reduce the time and cost associated with data curation. SciBite automates the process of cleaning and normalising data, giving the foundation for downstream data discovery and integration activities. SciBite provides the high quality, contextualised data necessary for sophisticated tools such as machine learning natural language processing to be effective.

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