Automated Data Cleansing and Standardization

- **INTUITIVE** – Interactive sheet for fast and simple curation of data with terminology standards
- **AUTOMATED** – Reproducible annotations using vocabs or public/private ontologies
- **REPEATABLE** – Store annotations and share rules to reduce the time to curate new data.

SciBite Workbench serves to support organizations that are adopting a FAIR (Findable, Accessible, Interoperable, Reusable) approach to data management. A critical component to making data more FAIR is to enable data interoperability through aligning data to shared terminology and ontological standards.

Cleaning datasets and aligning them to standards can be an arduous task, often requiring specific expertise in both the subject domain and the available standards. Scientific data curators are often the unsung heroes within an organization, without whom, much of the downstream integration and analysis would not be possible. Workbench aims to support scientists by streamlining the curation process through a simple and intuitive user interface.

An important aspect of Workbench is the ability to reuse and repeat previously seen curation, saving time, and allowing teams to focus on defining and expanding the interoperability of their data.

In Workbench, data is made **findable** for both humans and computers; **accessible** to authorized users; **interoperable** with other applications or workflows; and **reusable** alongside other data and in downstream processes (see figure 2).
Accurately curate complex data
Annotate data from selected columns with your chosen vocabulary or ontology, including SciBite’s VOCabs – a library of manually curated vocabularies enriched with >20 million synonyms.

Dealing with ‘messy’ data
Workbench includes fuzzy-matching logic to handle variations in spelling and typographic errors. For data riddled with internal codes or proprietary terms, you can use Workbench to map them to your chosen ontology terms or vocabulary with ‘annotation rules’ eliminating the error-prone and monotonous process of editing. Workbench promotes replicability by storing these annotation rules, which can be re-run during subsequent data processing jobs (see Figure 3).

UI vs API. Why not use both?
Workbench comes with a powerful REST API that automates the same core functions from the user interface (UI) in a typical workflow, quickly processing a much larger volume of data with ease. All API results can be accessed via the UI, which is ideal for spot-checking and quality control.

Easy data sharing and export
Workbench has built-in sharing functionalities for collaborative curation projects. As a data owner, you can create a group where you can invite other colleagues to view or edit your annotations, see Figure 4. Annotated data can be exported in excel for use with many 3rd party tools.

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.