Image: © Freepik

Discovery Decisions

Software creators play a vital role in helping project leaders and decision-makers have direct access to data from within the tools they use for their analyses. Therefore, collaboration is a key component when ensuring data is managed correctly

Dr Matthew Segall and Dr Chris Leeding at Optibrium

The discovery of new drugs is driven by data. The data generated on compounds' activities – in combination with absorption, distribution, metabolism and elimination (ADME), physicochemical and safety properties – enable selection of compounds for progression to further study. These data are also used to identify relationships between the structures of compounds and their observed activities and properties – structure-activity relationships (SAR) – guiding the design of improved compounds against a project's objectives.
In accessing these data, time is of the essence; selecting compounds based on incomplete or out-of-date information or designing new compounds without the latest results can lead to wasted time and effort through redundant experiments being performed and poor compounds being synthesised.

Managing Data

Large pharma companies use sophisticated laboratory information management platforms to capture information and store the results in large databases or data 'warehouses'. Entire departments are devoted to ensuring the coherence of these data because of their value to the company. Smaller organisations often do not have this luxury, and, in the worst cases, may have data from a variety of sources stored in individual files, in different locations and maintained by a range of scientists (often for their personal use). In the latter case, how much value is lost when data, often generated at high cost, are not available to the decision-makers in a project?

Even when data are consistently collected, stored and managed centrally, retrieving relevant information in a convenient form for analysis can sometimes be difficult. The raw data have limited value unless knowledge can be extracted with which to guide the selection and design of compounds. Many software platforms are used to visualise and model compound data to identify high-quality compounds and understand SAR to inform the next steps in optimisation. All of these platforms can read standard file formats, such as comma-separated value

> **"Data for a project are not static; projects move quickly, and, after completing an analysis, new assays will be run and fresh compounds synthesised and tested"**

or structure definition files, but, despite this, the output from data management platforms often cannot be read directly. Examples of the reasons for this include multiple values being stored within the same field in a file, compound structures being output separately from their associated data and different data for the same compound being split across multiple entries in a file or across multiple files. A lot of time can be wasted reformatting or pre-processing information so that they can be imported for analysis. Indeed, frustrations with this can even lead to data being ignored or software not being used to its full potential, wasting yet more time and resources.

Accessing Data

Most drug discovery project scientists are not informaticians and certainly do not want and should not need to understand the intricacies of relational databases, flat file formats and pre-processing of data. This should all happen behind the scenes. Project leaders and decision-makers should have direct, intuitive access to their data from within the tools they use for their analyses. This would enable scientists to spend more time applying their own expertise and adding value to the projects on which they are working. However, creating a solution for this presents challenges for software creators.

First, a query interface must be user-friendly, making it easy to define both the search criteria and data to return for analysis, and be independent of the way they are actually stored and retrieved. Few people want to concern themselves with learning and using database languages such as Structured Query Language (SQL). Additionally, even with an easy-to-use interface, creating a new query from scratch is not always necessary; often, the same query, maybe with small modifications, will be run many times. Sharing commonly used queries – perhaps defined by a power user – across a project team can save everyone a lot of time. It is also important, therefore, to be able to save, share, edit and run pre-defined queries.

Of course, drug discovery data come in many forms. Compound structure information is key to medicinal chemists, so it is crucial to be able to search for specific structures or substructures, for example defining a chemical series. Biological data may be numerical, eg IC_{50} values, percent inhibition, ADME properties or *in vivo* efficacy or pharmacokinetic parameters, but, in some cases, results may be categorical, eg high/low. Text fields, such as compound IDs and quality control comments, and dates for assay results or compound registration may be used to track progress in a project. Therefore, it is important to be able to define search criteria based on any of these types of data and combine these with logical 'or' and 'and' to find only the most relevant results.

Compound-related data have even greater complexity, in that the same experiments are often repeated multiple times for the same compound and the results aggregated to give an average value. However, these experiments may be performed on different salt forms of a compound or batches synthesised at different times or stored in separate locations. Therefore, scientists may wish to aggregate all data from replicates of an experiment performed on a compound or only for a specific salt form or batch to look for differences between the batches or salts. If they spot an unusual result, it is also valuable to 'drill down' to the underlying measurements to identify any outliers

Image: © Freepik

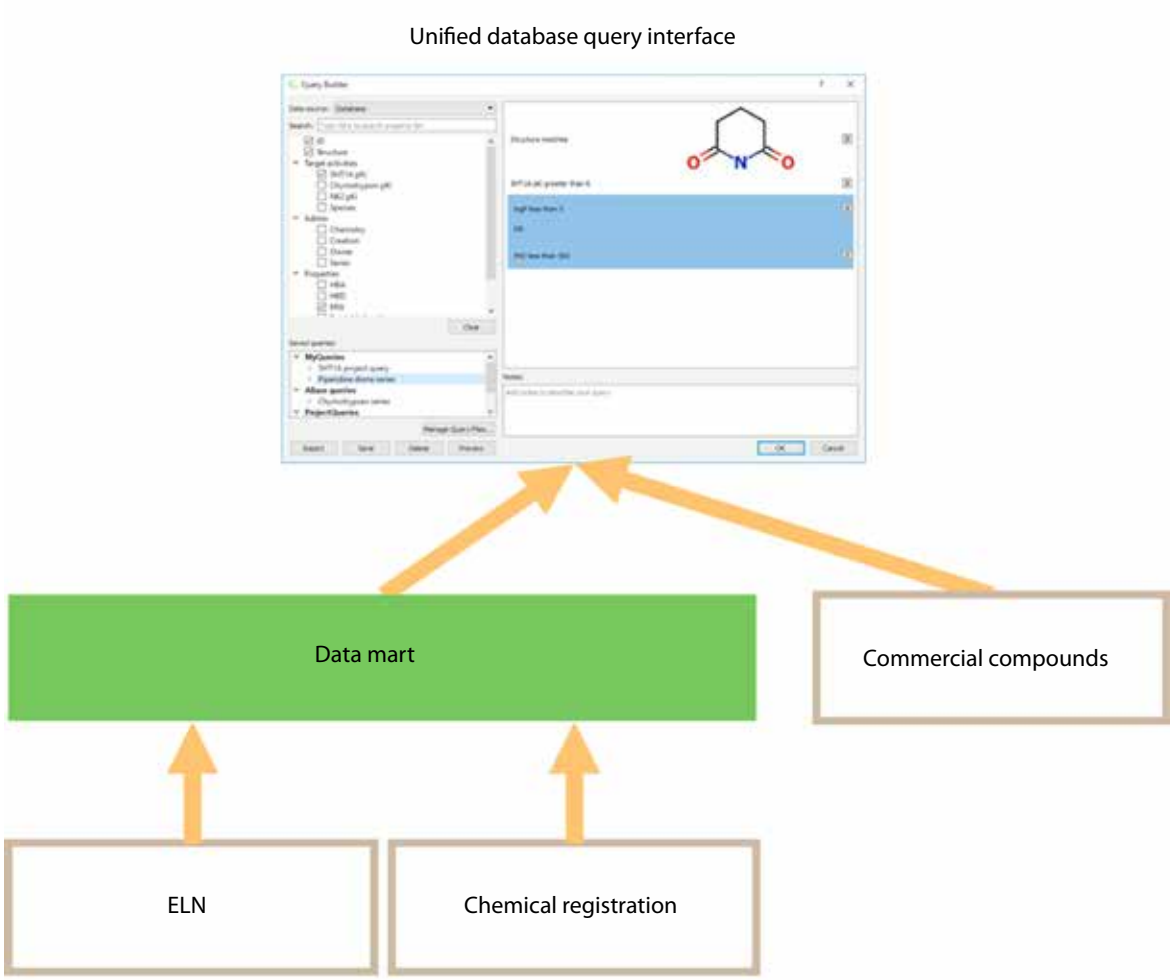

Figure 1: An illustration of the data architecture implemented at Zoetis to enable discovery scientists to seamlessly access biological and compound-related data from in-house and external sources through a single user-friendly interface, revealed by the technique and the close similarity of the two particles

that may be skewing the aggregated result or view metadata such as curve fits or quality control notes.

What is more, data for a project are not static; projects move quickly, and, after completing an analysis, new assays will be run and fresh compounds synthesised and tested. One approach to deal with this is to enable the results of a query to be refreshed, updating any analyses and highlighting any new data to draw attention to potentially important changes. This enables scientists to keep abreast of the progress of their project without the need to repeat their analyses.

Finally, even if in-house data has a single source, eg a data warehouse or 'mart', project scientists may wish to search other sources, such as external or public-domain databases. For example, it may be useful to search commercially available compounds and compare those with an in-house collection to find compounds for purchase in order to expand SAR around a hit or lead series. Enabling this through a common user interface makes combining data from different sources easier.

Case Study

Zoetis, a global animal health company, required a unified interface to their biological and compound data from their electronic lab notebook (ELN) and compound registration system, easily enabling access for their scientists. Their chemists wished to be able to search additional data sources,

"In whatever system the data are stored, it is important that it is linked seamlessly to the software used for visualisation, analysis and design of compounds"

such as a database of commercially available compounds. Figure 1 illustrates the architecture that was applied to achieve a seamless user experience.

An intermediate data mart was implemented to provide a single source for in-house biological and compound data, aggregated by 'concept' (parent compound), salt and lot. A unified database query interface was developed that provided user-friendly access to data in the data mart and additional, separate sources.

The result enabled Zoetis' scientists to seamlessly access information from multiple sources for visualisation, analysis and modelling to guide their discovery projects.

Looking Ahead

A data source does not need to be a sophisticated data warehouse; it may be a simple flat file, carefully curated and kept up-to-date. However, in whatever system the data are stored, it is important that it is linked seamlessly to the software used for visualisation, analysis and design of compounds.

Only then can the true value of the data be realised to guide decisions and move a drug discovery project quickly to a successful outcome.

About the authors

Dr Matthew Segall has an MSc in computation from the University of Oxford, UK, and a PhD in theoretical physics from the University of Cambridge, UK. At Camitro, ArQule and Inpharmatica, Matt led a team developing predictive ADME models and intuitive decision-support tools for drug discovery. In 2006, he became responsible for management of Inpharmatica's ADME business, including experimental ADME services and the StarDrop software platform. Following acquisition of Inpharmatica, Matt became Senior Director of BioFocus DPI's ADMET division and, in 2009, founded Optibrium, which develops software for small molecule design, optimisation and data analysis.
Email: matt.segall@optibrium.com

Dr Chris Leeding has a PhD in chemistry from King's College, London, UK and his career has focused on scientific software development. Chris has worked on the development of StarDrop since 2006, most recently as Optibrium's Director of Product Development, and taken a lead role in integrating StarDrop with customers' database and informatics infrastructure.
Email: chris.leeding@optibrium.com