Case Study

How Data Management Helped Increase Efficiency Through Computer-Assisted Structure Identification

The Problem – Scattered and Inaccessible Data

Just over a decade ago, a leading international tobacco company decided it was time to put its analytical data to better use. At the time, spectral data was being collected from a variety of instruments from different vendors for mixture analysis and component characterization. Results were on shared drives or the individual computers of 30-50 scientists. An immense amount of data was scattered and inaccessible to help with other projects.

> *There was no centralized database where everybody was able to review if a Spectrum had been recorded previously.*

Overview of Accelerated Identification of Components

A leading tobacco company realized major shortcomings in its data storage and structure identification process. Data was scattered all over their company and therefore unusable.

A centralized database containing reference spectral information, structures, and retention indices was created as part of the computer-assisted structure identification process.

More recent cloud-based deployment allows for faster and streamlined access to stored and available data. Computer-assisted structure identification (CASI) has improved the accuracy and speed of their structure identification.

The Goal–Centralize Data Access

With the goal of centralizing data access, a team was made responsible for setting up the infrastructure and tools necessary to build a database of chemical and spectral data. The aim was to leverage existing data and reduce the effort of manual structure identification of aerosol components in smoke from conventional tobacco products and e-cigarettes.

The Requirements

- Software with the capabilities to process and store GC/MS data, no matter what instrument or format the data was collected in
- The capability to store method data and other chemical information (e.g., retention index) with spectra
- Ability to add spectral and chemical information automatically on an ongoing basis for centralized search and access

Solution Deployment

MS Workbook Suite was chosen since it fulfilled most of the requirements "out of the box". Analysts created a centralized reference database of GC/MS spectral and related chemical information for compounds of interest. They built a complete record containing mass spectra, the structure, and a company identifier for every instrument, a variety of columns, and methodologies (e.g., volatile, non-polar, polar), removing this burden from scientists.

> *If you use the same instrument and the same method, the match factor for mass spectra [of GC/MS datasets] is usually very good. Having this information in a centralized place makes component identification very fast. Storing match factor with the spectral information means that we have built-in quality control. When spectral match is good but the retention factor is different we know we should look into it more.*

The Outcome

Ten years later, the team migrated their deployment to the cloud, in alignment with IT strategy. With centralized and streamlined databases, their 20 users and 2 expert chem-information power users can implement a function called 'export to NIST format'. When a NIST record for a spectrum from a specific instrument is required, the power users will search for the spectrum and export it in a NIST format compatible with their workflows. Using computer-assisted structure identification (CASI) to match spectra with the reference library and retention index, they can identify compounds with improved accuracy and speed.

> *The database was created to speed up the identification of unknowns. In our CASI process we are matching spectra with this library as well as matching the retention index, which is really useful in component identification.*