

Data Management In the Spectral Laboratory

by
Marie Scandone & Deborah Kernan
Bio-Rad Laboratories, Inc., Informatics Division

In the laboratory, there is a strong initiative to bring diverse analytical disciplines into an integrated system so that information may be shared and utilised by an organisation. We have moved from the archiving and warehousing of data to tools that identify and evaluate information.

Today, data management is a necessity. In the simplest terms, it is identification, collection and effective management of intellectual assets within an organisation. Its main goal is to make existing data resources available to everyone in an organisation in order to meet the challenges of an ever-changing business environment.

Because valuable resources are used to generate information, it is logical that efforts should be made to capture and use all that information. Successful implementation of a data management system provides the means to maintain and improve the effectiveness of an organisation and increase productivity.

The continued success of any organisation may depend on how effectively data management is employed. With the rapid change of staff in every field and technology, critical information and expertise may be lost. Therefore, it is imperative that all information obtained by an organisation be maintained and utilised for the best results.

In managing analytical data, a system must allow the user to:

- Archive instrument data files
- Share instrument data files
- Compare instrument data files
- Search instrument data files
- Build analytical databases
- Access analytical databases
- Manage analytical databases
- Analyse analytical information
- Predict analytical information
- Report and communicate information

We can now capture existing data, save information and generate knowledge. Data systems are now designed to support decision-making and allow data to be stored, processed and managed. This approach is necessitated by the business need to effectively analyse all available relevant data as rapidly as possible to facilitate decision-making and to provide required information for regulatory compliance.

There has been great pressure for development beyond a LIMS system to bring diverse analytical information, consisting of multiple spectroscopic data types, to a variety of workgroups in an organisation. Traditionally, there has been a lack of common interface that has made using and sharing spectral data difficult.

A system must meet the needs of the analytical chemist as well as the spectroscopist, the medicinal chemist, or any researcher generating information in the laboratory.

In the past, there were limited options available for the analytical informatics environment. Many factors limited its implementation, one being the number of analytical techniques performed. Compounding this issue, within each technique, a number of operating systems are in use. The instruments used to produce the spectra can vary as well as the software used to generate the data.

Usually, one person is responsible for one operating technique so there is little need for a common environment. There are no standard data formats within the laboratory or even within a technique, and even then there may be difficulties in reading

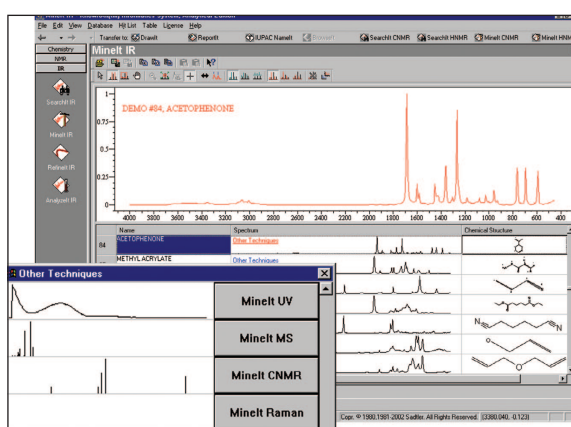


Figure 1. The KnowItAll® Informatics System offers software tools and data for multiple analytical techniques.

different data types. In numerous cases where an instrument vendor has products in several analytical areas, no attempt has been made, even by the vendor, to facilitate data combination or exchange within their own product lines. Although there have been, and still are, systems that can work with multiple types of data, this lack of a common interface for working with multiple spectroscopic data types has made using and sharing spectroscopic data difficult. The systems were not fully integrated or could not work with multiple types of data.

However, there has been strong impetus, especially from the pharmaceutical industry, to share information from diverse analytical disciplines. This need has arisen from the realisation that escalating costs for drug development dictate a "fail early, fail often" paradigm that reduces the number of poor or unlikely lead candidates as early in the drug R & D pipeline as possible. Some companies have come to realise that parallel efforts in analytical chemistry, for instance the use of NMR and mass spectrometry, could have yielded earlier and more cost-effective decisions on drug candidates if these data types could have been combined earlier into a single data management system. As the amount of spectral data increases, so does the need for accessing, processing and examining that data.

Spectral Databases

An important factor of spectral data management is the spectral database itself. It may be internally generated or externally created. Whatever means is used to generate spectra, the goal is the production of high-quality analytical data through the use of measurements that are verifiable and accurate. A collection of spectra can afford a convenient and easy access to the spectra of compounds having the desired functional group which can be used to establish an identity or classification through empirical comparison, as well as become a permanent record for a substance.

The production of quality data is essential to establish a laboratory's credibility and reputation and to satisfy the requirements of the users of the laboratory's services. In addition, the recipients who make use of the information may lack all of the required expertise to judge the data quality itself, but rely upon the establishment of verifiable and detailed quality control procedures to ensure that data they receive is of known quality in making day-to-day decisions. Partnering with information professionals to facilitate data management within an organisation can leverage the expertise in the organisation.

The analytical market has changed drastically over the past twenty years, but there is still a need for reference databases. Scientists need analytical tools that are easy to use, reliable, and pertinent to their work so that they can be free to perform the more valuable and creative duties in the laboratory. These tools must be effective and competitively priced in order to enhance the laboratory's worth and their bottom line. High-quality, digital, spectral databases are important tools in any management system. They are useful in identification, classification, verification and structural determination. A number of factors can influence the final product so all data must be tracked. Database quality must be a priority from the start and a formal database quality program must be in place to ensure accurate, correct and reliable information.

Database Building

When using spectroscopic techniques, management of spectral data is a key element in research and effective problem-solving. With the ability to capture and manage spectral data, analytical chemists can evaluate the data and accelerate development of new products. Data management of spectral data is more than the storage of data, but usually data storage is the first step. A laboratory must have the ability to build a database from a number of different instruments with possibly different operating systems. It must be able to cross-reference that information to other analytical techniques. It must be able to store spectral data as well as molecular structure information. Equally essential is the ability to store instrument parameters, physical and chemical properties, images and documents or any information generated related to a file, and it must be archived in a manner that allows easy access. Chemical drawing tools, as well as a means of chemical annotation, are required especially when dealing with a large number of records. It would also be

helpful to build databases for all techniques, including IR, NMR, Mass Spectrometry, UV/Vis, GC, Raman, NIR, etc.

A hyperlink as a property field can add to the effectiveness of any database. A user should be able to add links to web pages and other files such as spreadsheets, MS Word documents, and other valuable resources central to the organisation structure. By allowing hyperlinks to be stored in any database, a user can bring the World Wide Web into the management system and extend the power of the data to be linked, not only to files on the Internet but also to files on a local computer, local network, or Intranet. In addition, through web links, a management system can mimic the way scientific experts think and integrate information by placing critical related information onto the desktop to allow quick access and integration between collected and archived spectral database and other resident knowledge that must be a part of the decision-making on a project.

Search Capabilities

Once spectral data is archived, there must be a mechanism to search the data. Various methods of searching should be available. The first search method is the identification search. This may be a barcode number, a name, a lot number or any unique identifier used to mark the spectrum. The next method of searching is the spectral search. This allows the comparison of similar data in a database, and when using external databases, provides a means to match or locate spectra. In applied spectroscopy, an exact duplication of two spectra is sometimes the exception rather than the rule.

Therefore, a search algorithm must consider peak location uncertainty when comparing spectra. The degree of this uncertainty is specified by the search tolerance parameter of the peak search. There should be access to more than one spectrum search algorithm since results may vary. Because the algorithms simply compare one data point to another, the first hit is not always the best.

A peak search provides a quick way to match points in a query spectrum with spectra in the database. It attempts to match each peak in an unknown spectrum with each peak in a reference spectrum. A peak search should allow the user to specify positions at which the searched peak table should not have a peak, specify the minimum number of peaks to match, specify the default peak tolerance setting, and choose to ignore intensities if desired.

A property search greatly expands the flexibility of using additional information collected on a chemical. A record may contain alphanumeric chemical and physical information. Relevant information should be part of the permanent record of a chemical compound. Once recorded, a management system can search and utilise all pertinent information.

Finally, if a chemical structure is defined, the ability should exist to search the database for exact matches or substructure matches. Therefore, the ability to edit and/or create chemical structures as well as import files in different formats is required.

Analysis Tools

Analysis is often done on materials that are not readily available in spectral databases. Once these substances are synthesised, and to advance the development process, identification, or at the very least, classification is essential. For example, interactive software is available to assist in the interpretation of infrared spectra. It contains a database of characteristic group frequency spectral ranges and intensities. The results will display all of the functional groups in the database that contain a peak which falls within the tolerance set for the peak selected.

The software will also list the type of chemical bond that causes the absorption, any other characteristic absorption bands of that functional group, normal relative intensity for each band of the specified functional group and the substructure for that functional group. It can provide clear and rapid verification and identification by providing information on functional groups. The knowledge base can not only aid experienced analysts in interpreting complex spectra, but it

Figure 2. Being able to discover trends is crucial when evaluating large amounts of data. Powerful tools for mining data is included in the KnowItAll® Informatics System.

can also be used to instruct novice IR spectroscopy users in the basics of spectral interpretation. This accomplishes the goal of disseminating information resources throughout the organisation using a common yet powerful tool.

Prediction

There are a number of software packages available today to predict NMR spectra. The purpose is to expand the knowledge of everyday chemists to predict results using NMR instrumentation beyond the expertise of a few skilled spectroscopists. Using algorithms with the ability to predict shift values plus using additivity rules that are refined with the use of a large database of peak values to create an expert system, prediction may be provide spectral information to the user. Employing large NMR spectral databases is the most popular approach and it makes use of assigned chemical structures. Structures are described by HOSE (Hierarchical Organisation of Spherical Environments) code. In the prediction, the software searches for matches between the HOSE codes of the model and the database of chemical shifts.

As described, most proton prediction software uses database similarity or rule-based approaches which only accounts for two properties of chemical structures. They usually examine hydrogen type and connectivity. However, a NMR spectrum is clearly dependent on the three dimensional structure of the molecule as well as its flexibility. Therefore, prediction that examines the effects of the stereochemistry, intramolecular interactions and solvent effects is preferred. By incorporating these variables into the mathematical models, chemical shifts can be predicted with greater accuracy. The resulting prediction and simulation taking the molecule's hydrogen type, bond connectivity, 3D structure, solvent effects and flexibility into account becomes truly reliable.

Chemical Structure-Drawing Tools

An essential component of any management system of spectral data is a comprehensive chemistry drawing and publishing software designed for the chemist who needs to produce professional reports, chemical structures, chemical reactions, lab experiment setups, chemical engineering diagrams, data tables, and more. Communication of all results is essential and chemical structures can be used to identify results.

The user should have the ability to draw any chemical structure with little difficulty. Ideally, it permits the user to store fully integrated chemical structure fields and use pre-designed tables and forms to enter chemical data and perform substructure searches.

This makes building databases easy and efficient. The user should also be able to create links between objects and easily enter and organise data. The software should also combine the most advanced structure drawing tools with a module that understands systematic nomenclature rules.

Conversely, the program should also allow the user to create high-quality structures by entering a molecular name. This permits the management of database structures, graphics, and information associated with the graphics, as well as the seamless creation of high-quality reports.

Multiple Techniques

Spectral data is a valuable tool in confirming the identity and quality of a compound. It is expensive to create spectral data, so it becomes a valuable asset of any company. It can be used in drug discovery, R & D and regulatory compliance. Laboratories contain a heterogeneous mix of instrumentation from a variety of manufacturers. Each may be using software unique to that vendor. The instrument was chosen on its merits to perform a particular analysis, not on its ability to share data files. Since the amount of information is increasing, accessing, processing and examining spectral data has become an interdepartmental need. Companies must retain research information on the products that they develop and manufacture. Management of the data can reduce the time and costs to bring new products to market. It can meet regulatory requirements for archiving and tracking results. It can improve laboratory workflow and convert data and information into knowledge. It can accelerate decision-making and help to properly evaluate leads and new substances before they advance in the development process. A fully integrated environment for analytical techniques allows a user to transfer information from workgroup to workgroup and permits researchers to have access to all information in one place. Laboratory instruments may come with spectral processing tools but normally concentrate on one analytical technique at a time. Multiple users can readily share data while each user processes the data on his particular instrument. Collaboration is aided when a software environment readily supports other stakeholders in an organisation who must create, view or access archived analytical data.

Conclusion

In response to this need, Bio-Rad Laboratories, Inc. developed the KnowItAll® Informatics System, as shown in Figure 1. This data management system provides the analytical informatics consumer with increasingly efficient solutions that combine the means necessary to build, analyse, and access analytical databases with the ability to create, manage, and communicate knowledge from those databases. The architecture has been designed to increase access to information and share that information with less effort. With the combined power of a high-quality data set and the KnowItAll environment, researchers can search by spectrum, peak, structure or property, access reference spectra, make predictions, build databases with spectra, structures and chemical information, and even cross-reference data with other analytical techniques, such as infrared, vapor phase IR, Raman, Near IR, ¹³C NMR, ¹H NMR, mass spectrometry, UV/Vis, GC, etc., as well as provide ADME/Tox predictions and generate high quality reports and laboratory forms.

Combined with a large collection of spectra prepared using IR, Raman, NMR and Mass Spectrometry analytical techniques, this unique software facilitates the use of multiple techniques in the laboratory. It has the ability to manage data from multiple sources and helps users keep track of various types of spectral data. More importantly, it facilitates the management of spectral data by using the power of the data and making it accessible to all users. It also provides the user with the means to do ADME/Tox prediction. The software allows the user to view the results of prediction models such as: Absorption Rate, Bioaccumulation, Bioavailability, Blood-Brain Barrier Permeability, Elimination Half-Life, First-Pass Metabolism, Immunotoxicity, Irritation, log D, log P, Metabolism, Metabolite Toxicity, Mutagenicity, Neurotoxicity, Plasma Protein Binding, pKa, Polar Surface Area, "Rule of Five" Violations, Sensitivity, Teratogenicity, Volume of Distribution, Water Solubility and more, as shown in Figure 2.

Management of data can assist in preserving valuable knowledge and identification of that knowledge can contribute to an organisation's success. With new and powerful informatics tools, companies can now accomplish the goal of enterprise-wide information access, integration and enhanced decision-making.

Acknowledgments

We would like to thank Gregory M. Banik, Ph.D. for his input.