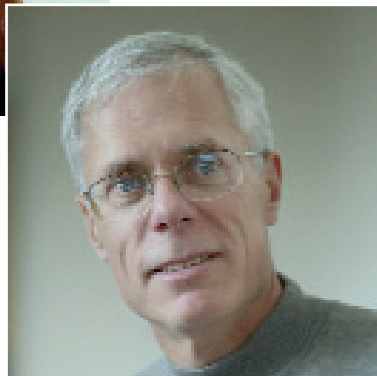


Guest Editorial

Preserving Analytical Data for the Long Term



Trish Meek^a and Jack Korpi,^b

^aThermo Electron Corp., Informatics, Altrincham, Cheshire, UK.

^bThermo Electron Corp., Informatics, Madison, Wisconsin, USA.

Long-term preservation of analytical data — including controlled access, content preservation, and storage and media management — has become a major concern of science-based organizations. If well addressed, an organization can reap the benefits of effective knowledge management while easing the pressure of regulatory compliance. Clearly, the management of chromatographic data is a significant part of this problem.

Chromatography instruments have become increasingly automated and now run significantly more complex analyses in much less time than ever before. Today ultra-fast gas chromatographs (GCs) are available that reduce runs that previously took 20 min to just 20 s. As a result, the volume of electronic data generated and stored in a laboratory today is many orders of magnitude more than ten years ago.

Both regulatory and internal pressures are compelling organizations to manage electronic information as they would any other business asset. The need for electronic record management spans the entire range of companies engaged in science and research, including pharmaceutical development and production, chemicals and petroleum, consumer products and many other industries. Although the interpretation and enforcement of 21 CFR Part 11 is currently a subject of much debate, rulings such as this are a further factor urging attention in this area.

A strategy of storing data from R&D projects for subsequent access is now seen as an important part of helping life science institutes and organizations make new discoveries, develop new products and bring them to market faster. Deploying long term archival systems can be costly but there are several ways to justify the investment:

- Reduce time and costs necessary to bring products to market, particularly for patented compounds and materials.
- Improve laboratory throughput, workflow and quality control.
- Retain important research and development data and convert it into “knowledge”.
- Gain efficiencies from shared expertise, instead of local islands of knowledge.

- New data-mining methods can now find relevant information in vast historical data collections.
- Meet regulatory requirements for long-term storage and traceable histories of data and results.

It may be tempting to adopt common desktop office applications and other commercial off-the-shelf (COTS) software applications for the archival of data. However, recent advances in chromatographic analysis have increased the sheer size and volume of data sets from analytical instruments to the point where the storage capabilities of these systems are simply inadequate. It is not unusual for a single R&D project to generate several terabytes of data.

The problem of preserving the physical existence of electronic records is only half the problem. Preservation of the inherent content for the long term is critical and, in fact, an even greater challenge. The longer the required retention period for an electronic record, the more time and technology advancement will jeopardize the ability to read that record. To preserve the ability to access actual information content, it must be extracted, translated or migrated to new storage formats. This is a significant problem for specialized, multiple industry market applications such as chromatography data systems (CDS). The large number and diversity of specialized software applications used by science-based industries means it is virtually impossible to provide access to the record's data content to all the individuals who require it. Some organizations have taken steps to alleviate this problem by standardizing on common applications and vendors. For example, rather than using workstation-based CDS, many companies are now standardizing on client/server CDS.

In addition to simplifying day-to-day system operation and management, use of a common CDS means a single chromatographic data file format. This is still merely a partial solution to data diversity when all the other types of data in the laboratory are brought into consideration.

Each specialized system for each technique generates electronic records that are unique to that vendor's proprietary software. This diversity becomes a significant obstacle to enabling distributed access to information within the organization. It is simply not feasible to provide every application needed to retrieve, compare and visualize the contents of the stored records to everyone who may need it. For this reason, simply storing the raw "files" and providing secure access is not an adequate solution.

Another issue is the continual cycle of technological advancement. A company may go through three or more instrument models — each with its own data acquisition and control system, and often with its own file format — over the ten to fifteen years of a drug development programme. Consider this against common record retention periods of 20, 25 or even 30 years, and practical access to and use of the original data system software becomes a highly unlikely proposition.

To illustrate how the problem of multiple data file types from multiple instruments and vendors affects a QA laboratory, consider the example of an HPLC method used to determine the amount of drug substance in pharmaceutical tablets. An unknown peak, previously not seen, appears in sample chromatograms. A comparison of IR spectra of the inactive excipient ingredients may determine if one of them might be the source of the impurity. The following tasks must be performed:

- Find a previous product lot that did not contain the unknown peak.
- Locate the IR spectra for the tablet excipients used in that lot.
- Compare the IR spectra from the current and previous lots and note any differences.
- Use HPLC to determine if the excipients with a different IR spectra contain the unknown peak.
- If a positive match is found, then the unknown peak must be identified, along with other product batches that contain the same impurity.

This process could be greatly simplified by having a common storage and retrieval system for all the analytical data (NMR, IR, MS, GC, etc.) associated with each product lot. The ideal system would enable searching for results that meet a variety of criteria and graphically viewing them without reliance on the software used to acquire them originally. In the instance of IR spectra, the latter capability is especially important if the results were created using instruments from different manufacturers.

Owing to the issues described, the requirement for industry to agree on an acceptable global standard for analytical data has never been greater. This development of such a standard has been a preoccupation of scientists, informatics vendors and standards bodies over recent years. One such proposed standard is a schema known as GAML (Generalized Analytical Markup Language), based on XML (eXtensible Markup Language).

XML is gaining considerable support as the basis for scientific file format standards. A number of recent high-profile initiatives have proposed XML for the handling and normalization of analytical data. A new ASTM subcommittee E13.15 "Analytical Data Management", comprised of representatives from instrumentation vendors and industry, has chosen as its first task to define an XML standard for analytical

instrument data. Conversion technology now exists to allow the creation of accurate and complete representations of scientific information based on XML, including complex instrument data records.

...the requirement for industry to agree on an acceptable global standard for analytical data has never been greater.

Storage and Media Management

Controlled-access storage and content preservation are, in fact, just two parts of a three-part requirement. The third is storage and media management.

Having a means by which to track and manage "classes" of electronic records and migrate them across media and platform as technology and business requirements change is an essential part of any long-term record-management strategy.

Long-term storage of electronic data requires balancing the costs of storage against the need for convenient access. An organization's archival strategy must take into account the inverse relationship between per Gigabyte costs of storage and access times. Further complicating this is the need to group logically related records onto the same physical media. For example, all records from a specific product or compound need to be placed on the same optical platter, making it easier to locate and access them in the future.

If companies are to truly leverage the value of historical analytical data as a business advantage, there is a need to deploy a long-term archival solution that genuinely offers ongoing access and reuse of information. The challenge is threefold: long-term content preservation, controlled and cost-effective access, and physical storage and media management.

Trish Meek is Data Systems Product Leader at Thermo Electron Corporation. She has a BS in Chemistry from Loyola University, New Orleans, USA. Trish joined Thru-Put Systems Inc. in 1999 in North America, which was later acquired by Thermo. She is currently responsible for product positioning and coordinating cross-Thermo development projects for Thermo's Atlas chromatography data system.

After many years' experience in instrumentation and laboratory informatics, **Jack Korpi** is currently Product Marketing Manager — Chromatography in Thermo Electron Corporation.