

Implications of 21 CFR Part 11 guidelines on the archival of analytical data

Adrian Fergus and Don Kuehl

The long-term archival of analytical data, including spectra and chromatograms, has been the subject of much attention and debate recently as the full implications of the introduction of 21 CFR Part 11 have been realized by organizations operating in the regulated environment. It has been suggested that compliance with Part 11 remains somewhat of a moving target as the industry continues to refine its interpretation of the rule, which became effective in August 1997. In order to help address this situation, the U.S. Department of Health and Human Sciences of the FDA issued a draft guidance document (docket number 00D-1539), published in the *Federal Register*, September 5, 2002.¹ The latest guideline focuses on how industry should maintain electronic records in compliance with Part 11. As a long-term approach, the agency has indicated support for a process of migrating records from one computing environment to another, provided measures are taken to control factors that might affect the reliability of records.

The document is intended primarily for regulated science-based organizations. However, it also offers assistance on compliance for FDA personnel, as well as computing software vendors serving pharmaceutical and other industries operating to GMP, GLP, etc. All of the interested parties see the opportunity provided by these latest guidelines to investigate new file formats and methods of migration that may offer compliance more seriously. This paper proposes a data model based on eXtensible Markup Language (XML) as a suitable format for the translation and migration of accurate and complete records as computer technology evolves.

The Agency defines two acceptable approaches to achieve compliance in handling the maintenance of electronic records for the duration of a record's lifespan (i.e., its retention period) (Figure 1):

1. *Time capsule.* This approach involves the preservation of the exact computing environment in which the data were acquired and processed. The document accedes that this approach is only viable as a short-term solution due to cost and technology advancement. This view reflects a major concern that has existed within science-based industry over recent years: that retrieval of electronic data can become reliant on the original computing software and hardware platform on which they were acquired. In fact, the problem is even more acute since each version of the software and hardware, at the point in time the data were

acquired, needs to be retained and, indeed, maintained in order to allow retrieval of the raw data.

2. *Data migration.* This approach involves the translation and migration of records through, if necessary, several successive computerized systems during a record's retention period. This effectively allows organizations to take advantage of computer technology as it evolves throughout the lifespan of the record. However, there is clear direction on what should be possible when handling the migrated records in their new state. To quote directly from the document, point 6.2.1.4 states: "In the migration approach, the new computer system should enable you to search, sort and process information in the migrated electronic record at least at the same level as what you could attain in the old system (even though the new system may employ

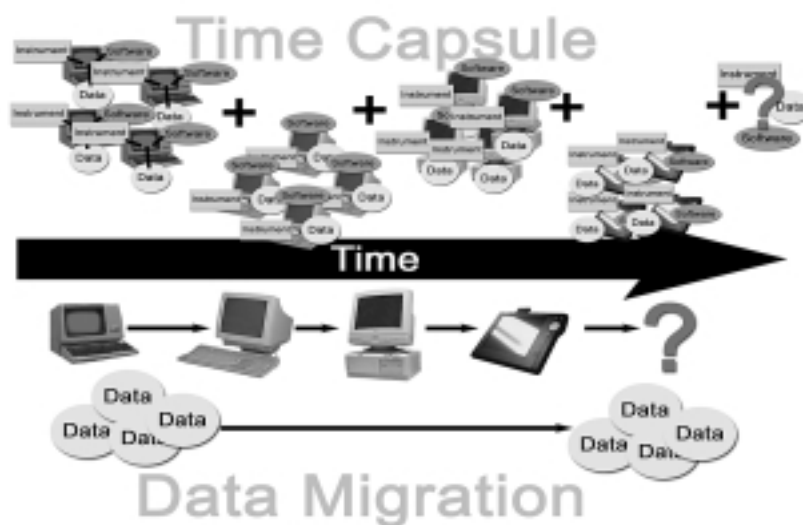


Figure 1 Time capsule approach to maintaining electronic records which preserves the exact computing environment including the computer and operating system as well as the vendor application software. This means the creation of individual software museums to enable long-term access to the data content. Conversely, the Data migration approach allows organizations to move forward over time with evolving technology.

different hardware and software).¹ This guideline appears to preclude the use of static images in favor of more dynamic representations.

The guidelines also advise the control of factors that affect a record's integrity, such as data encoded within the record, meta-data for an electronic record, and the process of extracting and presenting information in a readable form. For this approach, the Agency recommends the use of periodic testing to ensure that records remain readable.

The above descriptions merely summarize the guidance document as it refers to the long-term maintenance of electronic records. However, the protection of intellectual property within science-based organizations has been a concern of industry for many years. The need to implement data management systems and processes, which support compliance with GLP, GERM (Good Electronic Record Management), Part 11, and other important guidelines and regulations, added to the concern and pressure. This has been the catalyst for industry, standards groups, and informatics vendors to hasten efforts to develop an acceptable global standard for analytical data. One such proposed standard is a schema known as GAML (Generalized Analytical Markup Language), which is based on XML.

Unacceptable formats

While the regulators are slow to endorse a new standard for file formats for any data type, they are quick to define what is not acceptable, and this is useful guidance in the development of a new data model. One example is a section from the Drug GMP Report, which states: "Motise said the FDA thought that PDF file formats did not permit the processing of record information and thus would be problematic. PDF is a static format that does not allow reviewers to manipulate data to view subsets or generate analyses, tables, or graphs. PDF formats had been supported under earlier FDA policies because they were difficult to alter and were

widely used within government. However, the agency is now touting the Extensible Markup Language (XML), which is a more dynamic format."² The last sentence is noteworthy. Clearly, organizations solely utilizing static images as an approach to archive electronic records will not be in long-term compliance with 21 CFR Part 11.

Accessing analytical data for compliance

Before an attempt is made to propose a solution, it is essential for science-based industry to fully recognize the scale of the problem that may be caused through inaccessibility of data in the demonstration of compliance. While there are commercial benefits that could ultimately accrue from a common data format, regulatory compliance will be the key driver for change, and a common solution will be found that adequately satisfies the regulatory bodies. For example, consider the case of a proteomics researcher whose work is the subject of regulatory audit. The FDA inspector requests to see a specific piece of data on the particular protein on which he is working. The researcher knows that volumes of related different data exist within his organization, but has little confidence of being able to retrieve specific records. He is aware the inspector may ask to view and explore any related sample, spectra, annotated sequence, or 2-D gel image of the protein, which are held across multiple systems within his organization. It is only possible for the researcher to access these data by locating the archived work and loading onto his PC the original hardware, operating system, and software that acquired each instrument data type for each technique. This is due to the fact that each of the instruments for each technique has different data formats, even those from the same vendor. To exacerbate the problem, all of the formats are proprietary. It is also likely that all the hardware has to be maintained in-house and that

the expertise to support the operating system is also on hand. An alternative solution, albeit wholly unrealistic, to this *ad hoc* approach is to install every software package available on every computer in the company to allow all researchers to view the hundreds of formats. For the researcher to respond to audit requests with efficiency and confidence, he needs to have a global view of all related data.

The above example demonstrates the regulatory requirement for a common data format for all analytical data that will migrate across computing platforms, as proposed in the draft guidance. In one fell swoop, such a solution would take into account the evolution of instrumentation technology, storage media, and operating systems. Central access to analytical information, held in single common format, eases compliance.

Regulatory and industry support for XML

The attributes of XML as a basis for a data file standard have been well documented; XML is enjoying widespread acceptance both as a data interchange and storage format, as well as within the regulated industries. It is a public-domain, platform-neutral data-formatting standard that offers an application-independent way of representing data, however rich, using plain ASCII text. In order for XML to be effective in different industry sectors, rules for its use and structure must be documented to enable controlled and disciplined use in specific applications. To achieve this, a schema is used. This enforces numerical accuracy using standard ASCII characters.

According to a December 2000 survey by Silico Research (London, U.K.) that focused on pharmaceutical R&D, over 75% of the executives who responded indicated that they were planning to deploy XML as part of their R&D strategy or commercial product. Nearly all said they expect to be using XML by 2003. Regulatory support for XML is evident in a number of key areas. For example, the FDA is cur-

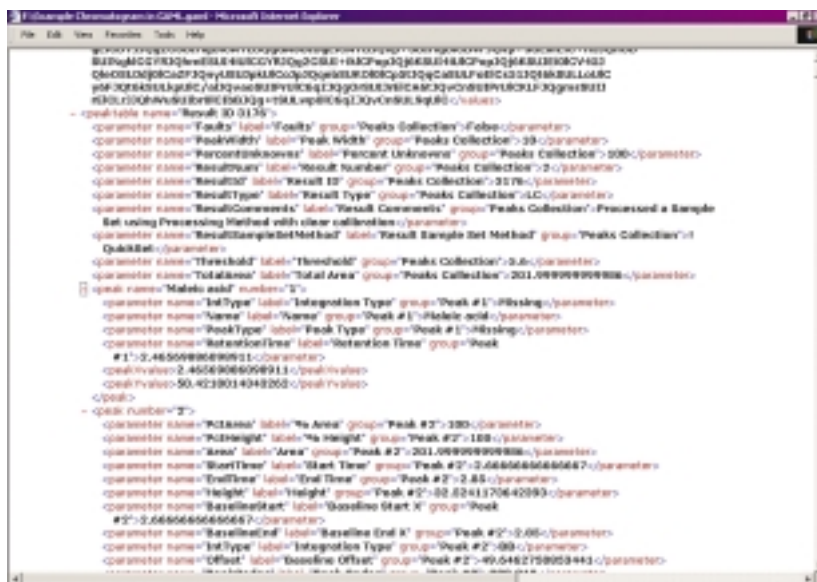


Figure 2 An example of part of a GAML file of chromatographic data. The file demonstrates the human readable quality of XML, on which GAML is based. Even to a chemist unaccustomed to XML, data can be easily interpreted and therefore represented graphically in the future.

rently proposing the use of XML in the development of the Cumulative Table of Contents (CTOC) in Investigational New Drug Applications (INDAs). XML has also been suggested as a format for Accelerated New Drug Applications (ANDAs), while complete XML schemas have been proposed by the FDA for stability and electrocardiogram (ECG) study data.

The trend toward XML appears to be gaining considerable momentum. There have been a number of significant recent industry initiatives in which XML is being proposed for the handling and normalization of analytical data. ASTM E13.01 (Spectroscopy & Chromatography), which includes representatives from instrumentation vendors and science-based organizations, has established a working group to define new XML-based data standards. Its equivalent analytical data standard group, ASTM E01.25, is investigating a Web Services model for vendor-independent data access, whereby an XML data model will be adopted for the handling of analytical data. Convergence of E01.25 within E13.01 to form a single ASTM XML working standard group is awaiting ratification. The fact that the qualities of XML have been evaluated by both

bodies for some time suggests that formal adoption of XML as a way forward may be forthcoming. Perhaps of even greater significance, the International Union of Pure and Applied Chemistry (IUPAC) is evaluating the implementation of a data standard based on XML. IUPAC is well known as the guardian of the Joint Committee on Atomic and Molecular Physical Data (JCAMP) format, a format that was originally defined for simple molecular spectroscopy data; efforts have been made to extend it to cover all data types.

XML schema for analytical data

With this backdrop of industry and regulatory support for XML, an XML schema known as GAML has been designed and proposed for the normalization of analytical data (Figure 2). Its strength is in its nonspecificity to any one instrumental technique. Despite all the qualities of XML, there are some formidable challenges to be overcome in the design of a single schema to meet the requirements of each different kind of analytical data. While a great deal of effort has been expended toward making GAML a very versatile format, while at the same time keeping it

reasonably simple and extensible, it likely has limitations.

Conclusion

For many years, science-based industry has puzzled over ways to guarantee long-term access to its archives of historical analytical data, while there is considerable pressure to increase efficiency and productivity in industries such as pharmaceutical R&D and manufacturing. The sharing of data, as the results of past research, among members of often globally collaborative projects is seen as vital to remain competitive. Nevertheless, in spite of this realization, it will be the obligation of industry to comply with regulatory requirements that will lead to the adoption of a standard data format for analytical data. If a standard format, such as the XML-based GAML, receives acceptance by industry, it could also serve as the common format to allow the migration of electronic records across platforms, and thereby comply with the FDA's recent draft guidance on maintaining records for their entire retention period. Data would no longer rely on their original computing platform for retrieval. In the absence of such an accepted format, there is a real danger that researchers will continue to require literally hundreds of software packages at every computer workstation in the organization to view the hundreds of analytical data formats that exist.

References

1. U.S. Dept. of Health and Human Services. FDA. Guidance for Industry, 21 CFR Part 11; electronic records; electronic signatures, maintenance of electronic records. Doc. 00D-1539. Pt. 6.2.1.4. Federal Register 2002; 5 Sept. 67(172):56848-9.
2. FDAnews.com. Drug GMP report Dec 2001; 113:7

Mr. Fergus is Product Specialist, International Marketing, Thermo LabSystems, Altrincham, Cheshire, U.K.; tel.: +44 161 942 3000; fax: +44 161 942 3001; e-mail: adrian.fergus@thermo.com. Mr. Kuehl is President, Thermo Galactic, Salem, NH, U.S.A.