TONY DAVIES COLUMN

AnIMLs in from the wild—XML standards at LGC

Mike Ludlow,[a] Dan Hopkins[a] and A.N. Davies[b]
[a]LGC, The Heath Business Park, Runcorn, Cheshire WA7 4QX, UK
[b]Professor of Analytical Science, SERC, University of Glamorgan, UK; Director, Analytical Laboratory Informatics Solutions

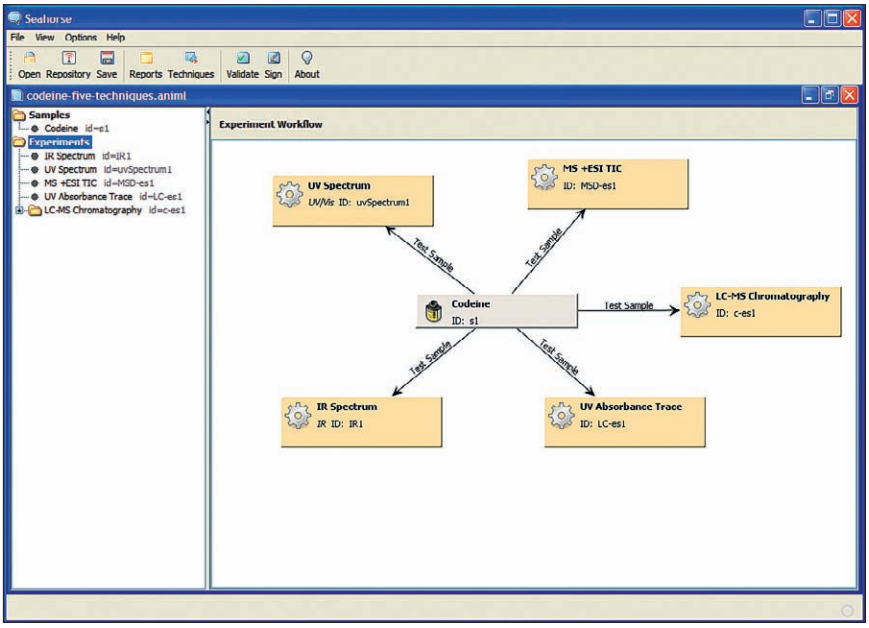
Introduction

I was very pleased when Mike Ludlow, a fellow campaigner from the EuroSpec project offered to report on the TopCombi Project within LGC. Not only does this project move the use of the draft AnIML XML data standard forward from purely archiving to a laboratory workflow technology. I was also please to see the involvement of Burkhard Schaefer, who has been a considerable driving force within the AnIML development team since his student days and time spent at NIST with Gary Kramer.

AnIML in "real labs"

Analytical Information Mark-up Language, better known as AnIML, has been around as a concept for a number of years, but how does an analytical chemist use it in the "real" lab? A team of R&D scientists at LGC has been finding out.

What is AnIML? AnIML is an emerging standard capable of capturing data from multiple analytical measurements, and is being developed within an ASTM working group. The AnIML data standard provides an open, universal XML-based format for analytical data and is suitable for use with many different analytical measurement techniques including chromatography, MS, UV, IR and NMR. AnIML files are designed to store analytical results, method information and sample data, and can contain the complete workflow description to allow repetition of an experiment. Additionally, the XML data format is now very widely used and is a strong candidate for long-term readability and archiving. It has strong support from industry, government and instrument suppliers, but to date its application within a working laboratory has not been well publicised. Funded by an EU project called TopCombi (Towards Optimised Chemical Processes and New Materials by Combinatorial Science), LGC has been working with Burkhard Schaefer of BSSN software to determine how the AnIML data standard could be used in practice in a real laboratory project using multiple measurements, and to understand the benefits and challenges associated with adoption of the standard.



Screen shot from the Seahorse software (BSSN) showing the experimental workflow.

The multidisciplinary nature of LGC's current scientific research, and advances in scientific instrumentation has highlighted the inadequacies in the traditional methods of recording, analysing and storing of time-sensitive scientific data. Data handling steps from high throughput measurements are a significant bottleneck in current research operations, with data being generated at multiple stages of the experiment workflow, on many different types of specialist laboratory equipment, and in many different formats. Vendor-independent, electronic data standards would have a hugely beneficial impact on both the productivity and quality of their R&D, facilitating the use of databases, enabling the uptake of generic data manipulation software and allowing data to be stored in a single readable format for extended periods of time.

In order to allow a step-change in the efficiency of data handling for their R&D labs, it is desirable for data from experiments to be generated and stored in a

single format. Whilst workflow-driven tools and automation can take care of file transformations to alternative formats, reducing manual interaction on data, standardisation would significantly simplify the required data handling steps, increase the potential use of data (e.g. through centralised storage in intelligently structured databases) and ensure long-term readability and access. In an age where commercial research is increasingly expected to have proper documentation of experiments and long term archiving of electronic data for 20 years or more, this aspect is becoming more and more important.

The team at LGC spent a considerable amount of time identifying a suitable analytical experiment to represent using AnIML. A model workflow was chosen covering multiple techniques including LC-MS, UV and IR. The goal was to collect sufficient detail of all the data for input into an AnIML file, to enable a third party to replicate precisely the experiments.

In order for LGC to evaluate the feasibility of the AnIML data format from their perspective and their customers, real data was produced for a typical reference material. Where feasible, data was generated directly from analytical instrument software interfaces, and BSSN software developed translation tools to convert this data into the AnIML format.

Adapting AnIML to meet LGC's needs

A significant amount of effort was focused on adapting the draft AnIML technique definition to allow all the necessary information to be captured for the LC-MS data. Extracting the experimental data and results from the multiple file formats generated and stored within the software was technically challenging. Each LC-MS run is captured in a folder on the file system, and the folder contains multiple XML and binary files. It was established that the XML files contain sample information, a device description, instrument settings, limited method information and various other configuration options. The actual method definition and the measured results are stored in binary files which are not documented. Parsing of the XML data files appeared to be feasible, so it was possible to extract many fields using this technique. However, accessing the result data proved to be more difficult. It was discovered that the current instrument software supports exporting the raw spectra to the mzData XML format, which is well documented, so it was possible to read and extract the raw data this way.

To automate the extraction of the spectra from the mzData format a prototype parser was implemented. This tool traverses the source document sequentially, looking for data elements that describe a mass spectrum. When a spectrum is found the data is extracted. Since the spectra are binary base-64-encoded, they need to be decoded and stored as an AnIML SeriesSet which can be used later. Some data that could not be generated in an output file from the instrument software was entered manually into the AnIML experiment files. To date the AnIML "technique definition", developed by the ASTM working groups, for LC-MS is in its infancy, and this work will be useful for future development of the standard.

Data volumes

The large volume of data produced by an LC-MS experiment presented an additional challenge when attempting to extract the raw data as the newer, accurate mass spectrometer instruments can create up to 1 GB per hour of operation. This makes it impossible to keep the whole data file in memory, so to overcome this issue a streaming parsing approach had to be used.

So what did LGC learn from this study and what is the future of AnIML? This work has demonstrated that AnIML is a viable format for accurately recording laboratory workflows and results from complex laboratory instruments. There were a number of challenges to overcome such as the lack of a fully defined AnIML Technique Definition for LC-MS. However, over time the number of techniques and their corresponding Technique Definitions will increase, and this work should be useful when the working group is formed to develop this element of the standard for LC-MS.

LGC, along with other scientific organisations, need the ability to share and preserve data in the long term without relying on expensive vendor specific software, and AnIML being human readable in an XML text based format will allow this. The key to AnIML's success lies with this ability.

Conclusions

The work conducted in TopCombi on e-Standards and the validation of the AnIML data standard has gained significant interest both from TopCombi partners and externally. A web-conference dissemination event organised by LGC and BSSN software (supported by ASTM) was held in November 2009 and was very well attended by 17 organisations, including those from the international pharmaceutical industry, software suppliers, national government agencies and academic groups. In the early stages of AnIML adoption, data translation tools will be required to generate AnIML files from various instruments and other software platforms. However, in the longer term it is anticipated that widespread adoption will lead instrument and software manufacturers to automate the generation of output files in the AnIML format and enable AnIML files to be uploaded into the systems from other sources. It is clear that AnIML has a bright future and looks set to expand from its roots in analytical chemistry to other scientific fields.

For further information about AnIML, BSSN, Topcombi and LGC see the weblinks below.

Comment

Interestingly, even though the adoption of these standards is in an early stage they are already intimately involved, as an essential key technology, ensuring continuing regulatory compliance during re-structuring within the pharmaceutical sector.—TD

Further reading

www.bssn-software.de/
www.lgc.co.uk/
www.topcombi.org/
animl.sourceforge.net/