Open Standard for LC/MS Data, Methods, and Results

David W. Peterson, Heiko Fessenmayr, Agilent Technologies, Inc.
Companies and scientists faced with long term data archival and recall requirements must surmount obstacles imposed by changing computational hardware, software, and proprietary vendor data formats. The Pharmaceutical Industry’s IQ Consortium has sponsored the Allotrope Foundation (www.allotrope.org) to develop and implement a fully open information standard for the analytical laboratory. Unlike earlier data format standards Allotrope’s standards encompass data, methodology, results and extensive sample information. Allotrope member companies from the pharmaceutical industry and Partner Network members (vendors) have engaged in this effort since 2012. The key concepts behind the Allotrope Data File (ADF) standards employ some of the most recent developments in computer science for organizing complex systems.

- Java and .NET class libraries to read and write the Allotrope standard formats from Microsoft operating systems and any Java virtual machine.
- Use of Web Ontology Language (OWL) www.w3.org to facilitate computer to computer exchange of information and relationships.
- A standard vocabulary of terms, definitions, units of measure to provide consistency across different instruments and vendors.
- Resource Description Framework (RDF) triples to describe “subject-predicate-object” relationships.
- Rigorous definitions for measurement quantities and units of measure - http://www.qudt.org/
- HDF5 Virtual file system from www.hdfgroup.org supported on multiple operating systems.
- ADFs are fully extensible for future applications, company or project specific metadata as well as vendor specific items.
- CMAPs Tool to establish visual representations of sample-predicate-object relationships for discussion and communication to knowledge engineers - http://cmap.ihmc.us/
- Numerous programming development tools to support the technologies using to create, write, and read ADFs.

The necessity for these technologies and rigorous development of the ontology and taxonomy becomes apparent when we realize how our definition of a SIM ion in our single quadrupole instrument has changed from a binary format (ChemStation) to an INI file format (ChemStation and MassHunter) and more recently to an XML format (OpenLAB C). The most common INI format represents SIM ions in a data acquisition method in the following manner.

Ion 1=200.00,100

This terse format does not clearly indicate which number represents the SIM ion and which number is the dwell time. Furthermore, the unit of measure for the dwell time is unstated. The proposed RDF data description being developed is shown below in an abbreviated format.

```
Run-1- Signal-SIM-Ion a «af:c:condition»;
 «af-x:has property» «af-x:m/z»;
 qudt:numericValue "251"^^xsd:double ;
 qudt:unit qudt-unit:MassPerCharge;

Run-1- Signal-ActualDwell a «af:c:condition»;
 «af-x:has property» «af-x:event duration»;
 qudt:numericValue "420.0"^^xsd:double ;
 qudt:unit qudt-unit:MilliSeconds;
```
Results and Discussion

Given the complexity and verbosity of the RDF triples which are intended for computer-to-computer interchange, CMAPs are used as a tool to visualize and communicate the relationships of mass spectrometers within the Allotrope teams and to the knowledge engineers and programmers. The corresponding CMAP for a SIM ion for LC/MS conveys the essence of the SIM ion and its attributes.

![CMAP diagram]

While the CMAPs are useful for the instrument parameters and results that populate the Data Description, the raw data is stored in the Data Cubes. In the case of UV detectors that are sampled at a constant frequency, a simple array in the Data Cube is written. In the case of mass spectral data, there are many types of data and the data varies based on the capability of the instrument. We chose to investigate storage for single quadrupole scanning data. When peak detection is employed, vendors write only the detected peaks into the raw data file resulting in scans that may have no detected masses or hundreds of detected masses. Initial experiments attempted to store each mass spectrum into an individual data cube as a two-dimensional array. When this experiment yielded undesirable results, three arrays were used in a single data cube.

<table>
<thead>
<tr>
<th>Storage</th>
<th>Mbytes</th>
<th>Creation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>917 Scan raw File</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Separate Data Cubes</td>
<td>26</td>
<td>80 seconds</td>
</tr>
<tr>
<td>Single Data Cubes</td>
<td>6</td>
<td>&lt;1 second</td>
</tr>
</tbody>
</table>

Obviously, the overhead to create data cubes is high. When a single data cube is used, the compression algorithms of HDF5 may be employed.

The data organization for the single cube is shown below.

Functional Software Prototype

Ultimately the ADF must be written and read in an environment that is commercially available to the supporters of the Allotrope Foundation. To demonstrate this, a prototype software was developed that supports LC instruments and LC/MS single quadrupole instruments on the ChemStation Edition of OpenLAB. The prototype consists of two components. The first component, the ChemStation2ADF Converter, writes the ADF format with method, raw data, results, instrument traces and other metadata. Once the ADF is created, it is automatically uploaded to an OpenLAB Enterprise Content Management system (ECM) by the Scheduler. The second component, the ADF filter, reads the Data Description from the ADF and places the information into a relational database and which is immediately available to all users through the ECM search and retrieval mechanisms.

Future Activities
- Support other Mass Spectrometer types
- Include qualitative results
- Contribute to the ADF standard for MS
- Read ADFs produced by other vendors
Results and Discussion

Agilent LC-UV and LC/MS Single Quadrupole

Convert ChemStation LC data to ADF / Upload to OpenLAB ECM

1. Convert to ADF  →  2. Upload to ECM  →  3. Filter ADF data in ECM

Conclusions

The technologies developed and promoted by the Allotrope Foundation have been demonstrated to work and are in use at several Allotrope member companies.

Acknowledgements

The authors acknowledge the contributions and support of the Allotrope Foundation, its Members and Partners as well as several key consultants.

For Research Use Only. Not for use in diagnostic procedures.