NMR Adapter Update

- A collaboration between Boehringer Ingelheim & Pfizer

October, 2019
NMR Converter – Bruker to ADF

- There is fully-functional NMR converter to convert Bruker NMR data to ADF
  - Supports: TopSpin, XWIN-NMR, UXNMR, UXNMR/P data
- Jointly developed by Boehringer Ingelheim & Pfizer
- MestreLab is taking the BI-Pfizer converter and is carrying it forward.

Command-line or GUI mode!
Data Cubes & Data Package

- Raw Data (FID’s) and Processed Data (Spectra) in the Data Cubes
- Proprietary-format vendor files in the Data Package
Data Description

Using a monolithic full graph ontology to describe all NMR metadata is …

- Unrealistic
- Unachievable
- Unsustainable

- 13,589 possible unique metadata values as of July 2019 and this does not include up to thousands more unique metadata values related to peaks, assignments, etc.
- Varian/Agilent & JEOL metadata not yet included in the model. Most vendor-specific hardware & software metadata are irreconcilably incompatible
- Deprecated values are repurposed & new values are continually added; definitions, datatypes & relationships evolve over time
... this is further complicated by the fact that all NMR metadata are not created equal ...

- Cross-vendor: Approximately 80 values
- Vendor-specific: Approximately 13,500 values
- Other Company: Approximately 10 values
… so trying to force-fit a very large volume of diverse, fundamentally irreconcilable, evolving, vendor-specific metadata into a single monolithic NMR ontology does not seem like a good use of valuable resources.

What are we to do?

How can we deliver value today?
Long, long ago and (not so) far away....
What is Allotrope Creating?

File format for any technique or instrument

Standard File Format

.adf

<table>
<thead>
<tr>
<th>TMS0025674</th>
<th>NMR Characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Allotrope from the WayBack Machine (2014)

What is Allotrope Creating?

Standard vocabulary & structure for metadata

With the Metadata Repository

<table>
<thead>
<tr>
<th>Project</th>
<th>Test</th>
<th>Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF0012354</td>
<td>IR Fingerprinting</td>
<td>380 FTIR/-SN/145453</td>
</tr>
<tr>
<td>AF0012764</td>
<td>Bulk and Tapped Density</td>
<td>ASTM Sieve-SN/3452</td>
</tr>
<tr>
<td>AF0012989</td>
<td>NMR Characterization</td>
<td>AM500-SN/0034578</td>
</tr>
<tr>
<td>AF0013142</td>
<td>Bulk and Tapped Density</td>
<td>ASTM Sieve-SN/09783</td>
</tr>
<tr>
<td>AF0045674</td>
<td>NMR Characterization</td>
<td>DRX600-SN/10234567</td>
</tr>
<tr>
<td>AF0034558</td>
<td>IR Fingerprinting</td>
<td>IS10 FTIR/-SN/341980</td>
</tr>
</tbody>
</table>

© 2014 Allotrope Foundation

Allotrope Connect - October, 2019
Leaf Nodes in ADF

• We can use greatly simplified semantic constructs called “leaf nodes” to represent this metadata in a *standardized*, highly *structured* way so that the ambiguity is greatly reduced – *in effect delivering on the promise of the “Metadata Repository” of 2014.*

In this use case, we’re not interested in the inter-relationships (branches) between the metadata values (leaves)
• Each NMR experiment has thousands of instrument settings that are represented in the original data something like this:

\[
\text{Acquisition Time} = 5.3
\]

• What does “Acquisition Time” really mean?
• Is the “Acquisition Time” on one instrument the same thing as another?
• What are the units?
• What kind of a number is it?
• Is it a number at all, or is it really a text string that sometime has letters?
Anatomy of a Leaf Node (simplified view)

Acquisition Time = 5.3

Node URI

has_Definition

“Time the detector is enabled”

Unit URI

has_Unit

<http://qudt.org/vocab/unit#MicroSecond>

“uSec”

has_Symbol

has_Label

“Acquisition Time”

has_NumericValue

“5.3”

Allotrope Connect - October, 2019
Anatomy of a Leaf Node (simplified view)

Node URI

has_Definition

“Time the detector is enabled” ^^string

has_Unit

has_Label

“Acquisition Time” ^^string

has_NumericValue

“5.3” ^^decimal

Unit URI

<http://qudt.org/vocab/unit#MicroSecond>

has_Symbol

“uSec” ^^string

Non-Allotrope data usually has just this

<http://purl.boehringer-ingelheim.com/ontologies/NMR#bi-t-BIT_00005041001>

Allotrope Connect - October, 2019
What does this look like in a real NMR example?

We have a lot of metadata!

A very simple case (6 LN)
The Net Result

- Leaf nodes provide a structured, lightweight, Allotrope-compliant way to instantiate NMR metadata in the Data Description

  - This still allows us to use the Allotrope “semantics infrastructure” to automatically apply constraints to the allowed values in an ADF and to execute standardized querying & validation of the metadata via the Allotrope API’s

  - Leaf nodes can be implemented to yield complete, functional ADF’s in days compared to the months… years… it can take to develop comprehensive, usable ontologies

- By a judicious choice of where the terms come from (Allotrope, Vendor or Member Company) we can simultaneously and easily accommodate the different types of metadata that are present in real-world NMR experiments (Cross-vendor, Vendor-Specific and Member Company)
So tying this all together into a process to permit rapid creation of ADFs with a minimum of overhead … and delivering value at each step along the way …

**Phase 1:** Instantiate all metadata values in ADF via a controlled vocabulary and a Member Company (or other suitable placeholder) URI namespace using a leaf node construct for speed & ultimate flexibility.

- **Phase 2:** Promote all terms (except Member Company-specific ones) to a controlled vocabulary with a vendor-based URI namespace: retain simple leaf-node construct.

(99% of all terms will stop here)
Phase 3: Identify the small number of “universal” terms with community input; utilize community SME’s to draft consensus PrefLabels & ScopeNote definitions; Submit SME-curated universal terms to Allotrope governance for AFT reconciliation and ultimate promotion to an Allotrope namespace. Instantiate Allotrope-governed universal terms in ADF as leaf nodes.

Phase 4: Elaborate relationships between “universal” terms only as needed to support bona fide use cases. Focused ontologies are only created out of a clearly and cogently articulated need in order to deliver demonstrable and quantifiable value to end users. At all costs avoid unrewarded complexity and effort to create ontologies simply to support “what if…” use cases.
• Fill data cubes & data package; metadata literals in data description (DD)
  ✓ Value: All data + metadata in enterprise in single, consistent, vendor-independent format

• Upgrade DD literals to member-namespace leaf nodes (LN)
  ✓ Value: Use semantics infrastructure of Allotrope API’s to define, validate, query metadata

• Upgrade member-namespace LN to vendor-namespace LN
  ✓ Value: Vendors get full control over metadata that apply only to their hardware/software

• Upgrade “universal” vendor-namespace LN to “universal” Allotrope LN
  ✓ Value: Achieve vision of truly vendor-independent universal metadata (applies to all vendors)

• Build ontology to support use cases from “universal” Allotrope terms
  Value: Data mining & advanced analytics
Questions?