The Empower CDS Allotrope Community Project

2019 Fall Allotrope Connect Meeting

October 8, 2019
Who is in The Community

Pharmaceutical Companies
- Baxter
- Bayer
- Boehringer Ingelheim
- Genentech
- GSK
- Merck
- Pfizer

Vendor
- TetraScience
Why are we doing this?

- Why do we want a CDS ADF?
- Downstream analytics/integration: Spotfire, ACD, etc...
- Long-term archiving: No software/version dependences
- Vendor agnosticity: Mix and match CDS... Empower to Chromeleon to ChemStation...
- Method development: Use/reuse the data from 100+ runs
- Many things we haven’t even imagined yet will be possible once we able to access data freely...
What is in this Project?

Empower LAC/E Acquisition Servers
Empower Server

Windows Environment

Empower Client/SDK

Empower Connector
Empower/ADF Converter

ADF Conversion App

<table>
<thead>
<tr>
<th>Description</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 Injections for XX</td>
<td>Conversion In Progress (25%)</td>
</tr>
<tr>
<td>50 Injections for YY</td>
<td>Converted</td>
</tr>
<tr>
<td>23 Injections for ZZ</td>
<td>Converted</td>
</tr>
</tbody>
</table>
How does it work?

- **Scanning**: from 0.01-0.1s per injection
- **Processing and JSON Conversion**: from 0.3-1.5s per injection
- **ADF Conversion**: from 0.5-2s per injection
- **System/Project Audit Trail**: 1 hour per 500,000 entries
- **Message Center**: 15 minutes per 50,000 entries
What’s the purpose?
Two practical use-cases unlocked by the project

1. Backup/Long-term-storage

2. Scientific data retrieval
Backup/Long-term Storage
Scientific Data Retrieval
<table>
<thead>
<tr>
<th>Project</th>
<th>3 Injections</th>
<th>Project</th>
<th>4 Injections</th>
<th>Project</th>
<th>5 Injections</th>
<th>Project</th>
<th>6 Injections</th>
<th>Project</th>
<th>7 Injections</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLTestProj</td>
<td></td>
<td>CIA_Default</td>
<td></td>
<td>Defaults</td>
<td></td>
<td>Defaults_FAT</td>
<td></td>
<td>Demo</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NIF_SP_WP_101414A</td>
<td></td>
<td>NIF_WN_WP_101514C</td>
<td></td>
<td></td>
<td>1672</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1880</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1680</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1896</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1688</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1904</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1896</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1912</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1904</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1920</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1912</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1823</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1920</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1823</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1848</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1840</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1856</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1848</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1864</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1856</td>
<td>Last Converted 4/3/2019 3:41 PM</td>
</tr>
</tbody>
</table>
### User Request

#### Request History

<table>
<thead>
<tr>
<th>Request Date</th>
<th>Description</th>
<th>Completion Date</th>
<th>Status</th>
<th>Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/3/2019 4:16:53 PM</td>
<td>PQ sample set injections</td>
<td>Conversion In Progress</td>
<td></td>
<td>Open</td>
</tr>
</tbody>
</table>

#### File Tree

- ADFOutput
  - allotrope_user
    - 20190403163412
      - Defaults_FAT
        - PQ Sample Set_1103

### Quick Access

- Desktop
- Downloads
- Documents
- Pictures
- bin
- nolog
- nolog
- nolog

### This PC

- Empower Data
- 7 Items

©2019 Allotrope Foundation
Live Demo of a working ADF...
1050+ fields extracted from Empower and available for conversion to ADF

41 data fields are mapped and converted to ADF. Most are SHACL validated.
What this means

Archival format?
Enabling search/analytics?
Fields Include the following:

Peaks Pattern
Results[].{ AdjustedTotalArea: "2788322" }
Results[].{ resultId: "2622" }
Results[].Peaks[].{ Amount: "3733.949" }
Results[].Peaks[].{ Area: 778910 }
Results[].Peaks[].{ curveId: "2389" }
Results[].Peaks[].{ endHeight: 0.0 }
Results[].Peaks[].{ endTime: 3.325 }
Results[].Peaks[].{ name:"Propiophenone" }
Results[].Peaks[].{ width: 21.5 }
Results[].Peaks[].{ height: 104391.0 }
Results[].Peaks[].{ peakLabel: “3” }
Results[].Peaks[].{ Resolution: NO_EXAMPLE }
Results[].Peaks[].{ RetentionTime: 1.249 }
Results[].Peaks[].{ Selectivity: NO_EXAMPLE }
Results[].Peaks[].{ SignalToNoise: NO_EXAMPLE }
Results[].Peaks[].{ startHeight: 0.0 }
Results[].Peaks[].{ startTime: 1.167 }
Results[].Peaks[].{ USPSPlateCount: NO_EXAMPLE }
Results[].ProcessingMethods[].SysSuitParamters.{ TangentPercentUSPPlateCount: “61” }
Fields Include the following:

**Mobile Phase Pattern**
InstrumentMethod.Setups[].{ SolventA: “Water” }
InstrumentMethod.Setups[].{ SolventB: “Methanol” }

**Gradient Table Pattern**
InstrumentMethod.Setups[].{ PctA: “40” }
InstrumentMethod.Setups[].{ PctB: “60” }

**Injection Pattern**
{ ID: 1191 }
{ DateAcquired: “9/17/1997 5:37:56 PM EDT” }
{ InjectionVolume: 20.0 }
{ Label: “U0104” }
{ ProjectName: “A166033” }
{ SampleSetId: 1342 }

**Sample Pattern**
{ sampleName: “Blank” }
{ Altered: “Yes” }
{ Superseded: “No” }
Fields Include the following:

**Submitter Pattern**
{ acquiredBy: “test_user” }

**Software Agent Pattern**
Results[].{ SourceSoftwareInfo }

**Sequence Pattern**
SampleSet.{ OriginalSampleSetId }
SampleSet.{ SampleSetId }

**Chromatogram Detection Pattern**
Channels[].{ ChannelId: 1192 }
Channels[].{ Channel: “486” }
Channels[].{ SamplingRate: “2.00” }
InstrumentMethod[].Setups[].{ Wavelength: “254” }
InstrumentMethod[].Setups[].{ Channel1Offsets: NO_EXAMPLE }
InstrumentMethod[].Setups[].{ Channel1Bandwidth: NO_EXAMPLE }
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Pattern Definition</th>
<th>Mapping coverage (Empower graph vs Instance data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peaks</td>
<td>100%</td>
<td>90%</td>
</tr>
<tr>
<td>Injection</td>
<td>100%</td>
<td>80%</td>
</tr>
<tr>
<td>Sequence</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Data Processing</td>
<td>100%</td>
<td>40%</td>
</tr>
<tr>
<td>HPLC Run</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Sample</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Gradient Table</td>
<td>100%</td>
<td>90%</td>
</tr>
<tr>
<td>Mobile Phase</td>
<td>100%</td>
<td>90%</td>
</tr>
<tr>
<td>Chromatogram Detection</td>
<td>0%</td>
<td>N/A</td>
</tr>
<tr>
<td>HPLC System</td>
<td>0%</td>
<td>N/A</td>
</tr>
<tr>
<td>Baselines</td>
<td>0%</td>
<td>N/A</td>
</tr>
</tbody>
</table>
[  
  a
  sh:ValidationResult ;
  sh:focusNode _:b0 ;
  sh:resultMessage "Less than 1^^http://www.w3.org/2001/XMLSchema#integer values have shape 0eaf968141a7f38390cf6865c22ae96f";
  sh:resultPath af-x:AFX_0000544 ;
  sh:resultSeverity sh:Violation ;
  sh:sourceConstraintComponent sh:QualifiedMinCountConstraintComponent ;
  sh:sourceShape _:b1
]
Only SHACL validation missing

[  
a
sh:ValidationResult ;
sh:focusNode _:b0; Not sure what this refers to
sh:resultMessage "Less than 1^^http://www.w3.org/2001/XMLSchema#integer values have shape 0eaf968141a7f38390cf6865c22ae96f";
sh:resultPath af-x:AFX_0000544 ;
sh:resultSeverity sh:Violation ;
sh:sourceConstraintComponent sh:QualifiedMinCountConstraintComponent ;
sh:sourceShape _:b1 ]
Possible Next Steps
Idea 1
Leaf Nodes
All have the same pattern...
# Leaf nodes Pros/Cons

## Pros

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| **Simplicity** | Easier to use (less onions to peel)  
|   | Easier to implement |
| **Human Readable** | Scientists can read the data more easily and  
|   | potentially obtain value sooner |
| **Lossless** | Any data field can be captured, as it’s not  
|   | constrained to an ontology |
| **Quicker time to value** | Value can be realized sooner without building  
|   | complex data-abstractions or huge learning  
|   | curves for end-users |

## Cons

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| **More difficult to do data mining and advanced analytics** | The relationships between terms are not fully  
|   | defined (ex: harder to do joins and draw  
|   | inferences) |
| **Not significantly different than tag/value pairs** | We will need to come up with patterns for more  
|   | complex data structures (lists, nested objects,  
|   | etc.) |
Idea 2
Map more fields
41 data fields are mapped and converted to ADF

We can continue expanding this by:

1) Expanding beyond the LC-UV reference
2) Mapping against the GC model
Expanding beyond LC-UV reference
Local Identifier (AFR_0000919)
Results[].Peaks[].{ curveId }

Written Name (IAO_0000590)
Results[].Peaks[].{ name }

Peak Width at Baseline (AFR_0001264)
Results[].Peaks[].{ width }

Peak Height (AFR_0000948)
Results[].Peaks[].{ height }

Number of theoretical plates by tangent method (USP) (AFR_0001240)
Results[].Peaks[].{ USPPlateCount }
Results[].ProcessingMethods[].SysSuitParameters.{ TangentPercentUSPPlateCount }

Retention Time (AFR_0001089)
Results[].Peaks[].{ RetentionTime }

Peak Area (AFR_0001073)
Results[].Peaks[].{ area }

Peak Analyte Amount (AFR_0001073)
Results[].Peaks[].{ Amount }

Peak Start Time (AFR_0001073)
Results[].Peaks[].{ startTime }

Peak value at Start (AFR_0001179)
Results[].Peaks[].{ startHeight }

Peak End Time (AFR_0001180)
Results[].Peaks[].{ endTime }

Peak Value at End (AFR_0001073)
Results[].Peaks[].{ endHeight }

Selectivity (AFR_0001235)
Results[].Peaks[].{ Selectivity }

Signal to Noise Ratio (AFR_0001195)
Results[].Peaks[].{ SignalToNoise }

Chromatographic Peak Resolution (AFR_0001230)
Results[].Peaks[].{ Resolution }

+ Additional fields
Using the GC model
SampleWeight is described in the GC Sample materials class model (Note this is not specific at all to GC, the GC model is being used to augment the chromatography model with items that were not included in the initial release.)
Idea 3

Search for ADFs and Backups
CURRENT PROJECT

ADF App ➔ Backup Folder ➔ ADF FILES

Backup Files
CURRENT PROJECT

ADF App → Backup Folder

Backup Files

ADF FILES

DOWNSTREAM (FUTURE)

Triple Store
Backup File Storage
Backup Search Tool
Backup Visualizers
Thank You/Questions?