NMR Data Conversion Project: Multiple Data Formats to ADF

Allotrope Foundation
CASE STUDY

Goals

- Transform multiple legacy NMR data formats to a single long-term preservation format (ADF) that retains all of the information content and analysis potential of the original data.
- Demonstrate the feasibility of creating and using the ADF for an inherently complex datatype: from one- to three-dimensional data, Kb to near Gb sizes, sparse data and non-uniform sampling must all be supported.
- Straightforward extension to hyperdimensional NMR data (4D & up) of multi Gb file size must be possible.
- Demonstrate that very old legacy datasets with archaic architectures can be successfully converted to ADF without the need of a bespoke IT project or solution.

Challenges

- 14 different variants of legacy formats.
- Diversity in word size (24- & 32-bit words)
- Diversity in endianness (byte order). Big- little- and mixed-endian data.
- Diversity in data dimensionality - 1D, 2D, and 3D NMR datasets
- Solution must be readily scalable to a large number of datasets (10^4 up to 10^6 and beyond)
- Accomplish the conversion with a reasonable resource commitment

Results

- Created an extensible architecture that can be expanded to handle data from other NMR vendors and techniques (e.g. FTIR, FTMS, etc.)
- Created a workflow and the applications to implement it in order to automate conversion of existing NMR data to ADF
- Successful population of ADF Data Cubes and the Data Package with content from legacy NMR data.
- Population of the Data Description proved to be a challenge due to the nascent status of the NMR-specific Allotrope semantic components (taxonomies, ontologies and data models). This led to the development of a "refresh" concept to populate the Data Description once semantic components of sufficient maturity become available.
- Demonstrated bi-directional, byte-for-byte, data conversion to/from ADF with zero information loss.

Goals

Since the Allotrope Data Format (ADF) has progressed to an advanced stage, there is now a need to demonstrate its utility in practical applications by using it for real-world data. This study leverages a very diverse body of legacy NMR data and attempts to transform it en masse into ADF while determining any limitations to the process including age/type and inherent complexity of the native vendor-proprietary data. The understanding gained in developing and implementing this process will aid in further refinements of the ADF and/or workflow for data conversion.

Challenges

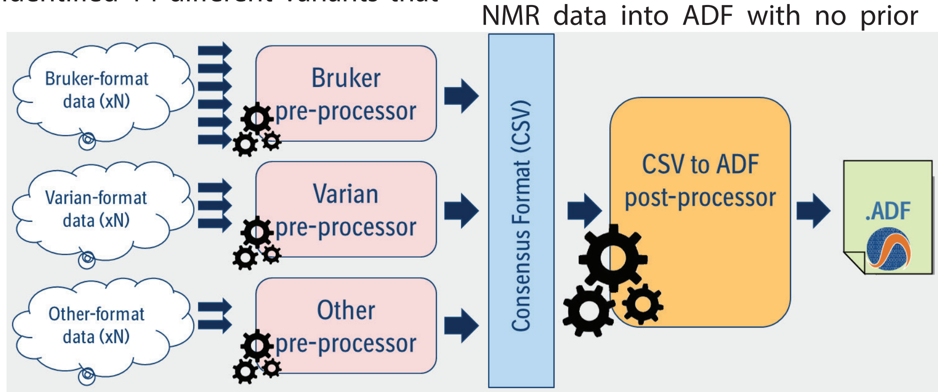
This NMR data project was intentionally limited in scope to one instrument vendor for simplicity. However, even with this simplification, the project still presented many challenges. Data from this one vendor (1D through 3D) were acquired over a period greater than 20 years and in this data we have identified 14 different variants that arose from different, continually evolving combinations of hardware and software. Amplifying this complexity is the concomitant evolution of the underlying IT technology – for example the transition from non-standard, proprietary 24-bit computer systems to commodity (32-bit and 64-bit) hardware. At the same time there was an evolution in CPU architectures, leading to differences in endianness (byte order; big-, little- and mixed) in our pool of legacy data.

A desired outcome is to identify any challenges to scalability of the ADF conversion process so that it will be eventually robust enough to be able to handle more than a million datasets with a reasonable time and resource commitment.

...data was acquired over a period greater than 20 years...

Results

An architecture, a resulting process, and the applications to implement it were successfully developed to convert this body of diverse legacy NMR data into ADF with no prior



http://www.allotrope.org
February 2017

knowledge or human intervention required. In the process, ADF Data Cubes (universal data container) and Data Packages (virtual file system) were successfully populated with NMR data from all 14 known variants. The third component of the ADF, the Data Description, proved to be more challenging to populate. In large part, this is due to the fluid state of NMR-related semantic components which are currently in an immature state but which are developing rapidly as part of a related initiative. While initially this appeared to be a roadblock to realize a full implementation of ADF, we have come to realize as a result of this effort that this is not as significant an impediment as we had first thought.

Developing a mechanism to refresh the Data Descriptions using the "latest" semantic components is useful to enable flexible data mining far into the future.

It is reasonable to expect that all of the Allotrope semantic components (including NMR-related) will continue to evolve over of time as the underlying science and vendor technology continues to advance.

Developing a mechanism to refresh the Data Descriptions using the "latest" semantic components is useful to enable flexible data mining far into the future. From this perspective, a key feature of any data conversion in the event that the Data Description cannot be completely populated is to assure that all vendor metadata is captured in the Data Package in its original, proprietary format so that it can be re-harvested in the future as needs and the Allotrope semantic components evolve. The concept of a "Data Description Refresh" application which facilitates this data harvest enables one to quickly convert data to ADF without the requirement of waiting for mature semantic components to become available. In the future, the ability to refresh the Data Descriptions allows newly selected metadata from proprietary vendor files stored in the Data Package to be added to existing ADF data descriptions as it is needed.

This demonstration project successfully converted a selection of NMR data encompassing all 14 known variants collected over 20 years. In addition, a reverse-conversion process was created to take NMR ADFs and regenerate, byte-for-byte, the original vendor format files for all 14 variants of the data. This back-conversion proceeds without any human intervention, or any prior knowledge of the original format. This assures that the original formats are re-creatable following ADF conversion for any potential future needs.

The development of taxonomies, ontologies and data models is enabled by working groups that are comprised of representatives representing both Member and Partner Network companies. For more information click here www.allotrope.org

About Allotrope Foundation

Allotrope Foundation is an international consortium of pharmaceutical & biopharmaceutical companies launched in 2012 with a common vision to develop innovative approaches for handling scientific data. Allotrope Foundation has developed a framework to capture and represent data generated by any analytical device in the laboratory in a standardized format, including more complete metadata related to each test and measurement event, expressed in a standardized vocabulary, which facilitates the exchange, utilization and integration of data beyond the boundaries of the originating instruments and laboratories.

This effort is fully funded by the members of Allotrope Foundation and is rapidly progressing on our common goals to improve data integrity, reduce wasted effort and allow us to realize the full value of our scientific data.

Allotrope Foundation
1500 K Street, NW
Washington, DC 20005
(202) 230-5855
http://www.allotrope.org