Whitepaper

ACD/LABS [ADVANCED CHEMISTRY DEVELOPMENT, INC.]

Standardization of Analytical Data: Best Practices

Richard Lee and Sanji Bhal

Improve Lab Productivity and Build a Foundation for AI-Powered Innovation

Why Standardize & Normalize Analytical Data?

Data is the cornerstone of innovation, driving insights from early research through product development and manufacturing.

What is data standardization?

Data standardization homogenizes data from different sources into a consistent *format*.

What is data normalization/harmonization?

Data normalization is broader than data format. Normalization translates data from various sources and systems into an agreed upon ontology (i.e. nouns, verbs, and adjectives that describe data and their relationships)

Organizations seek to standardize and normalize data to increase data accessibility and streamline integration of data from different systems and sources. Standardized data benefits R&D by enabling consistent data analysis, easier data exchange and collaboration, the ability to identify inconsistencies, and improve the quality of data being aggregated and used. It also allows organizations to leverage that data for AI/ML-powered innovation.

Analytical chemistry data, generated by various instruments and techniques (chromatography—CE, GC, HPLC, UHPLC, mass spectrometry [MS], nuclear magnetic resonance [NMR], infra-red [IR]/Raman/, ultraviolet [UV] spectroscopy, etc.), represents a wealth of information on the identity, properties, and behaviors of chemical compounds. Analytical instrument vendors, typically, create their own proprietary formats for data acquisition and handling. **The diversity of experimental techniques and proprietary instrument vendor formats results in data that is fragmented, incompatible, and difficult to integrate** into streamlined workflows. While most R&D organizations have successfully achieved some degree of digitalization of analytical workflows, data heterogeneity remains a major obstacle.

Heterogeneous analytical data:

- Hinders the assembly of interrelated datasets

- Impedes centralized data management and data accessibility
- Limits the use of valuable analytical data beyond its initial purpose

The latest approach to scientific discovery emphasizes the use of machine learning (ML) and artificial intelligence (AI) to uncover new insights by identifying patterns and correlations in massive datasets. **Data science (AI/ML) places a premium on well-curated, standardized data**. Analytical chemistry, with its diverse techniques and complex datasets, epitomizes the challenges and opportunities in aligning data with the computational demands of AI/ML.

Challenges of Analytical Data Standardization & Normalization

Homogenizing analytical data is hard.

Applications of analytical data are broad and varied. It is used every day in R&D labs to make decisions—both qualitative (e.g., "What is my sample?") and quantitative (e.g., "How much of each analyte is present?"); in the quality control of products and processes; as proof in regulatory submissions and/or responses, and more. Diverse data sources, constantly evolving instrument technologies, and complex regulatory requirements make achieving true interoperability difficult. Addressing legacy data while implementing future-proof standards for AI/ML adds further complexity.

The Current State of Analytical Data Standards

The need for handling analytical data in a homogeneous manner became apparent with the widespread use of computers and laboratory digitalization in the late 1980's and early 1990's. Historical efforts to standardize analytical data formats have evolved to facilitate interoperability. Figure 1 summarizes the most notable efforts over the decades.

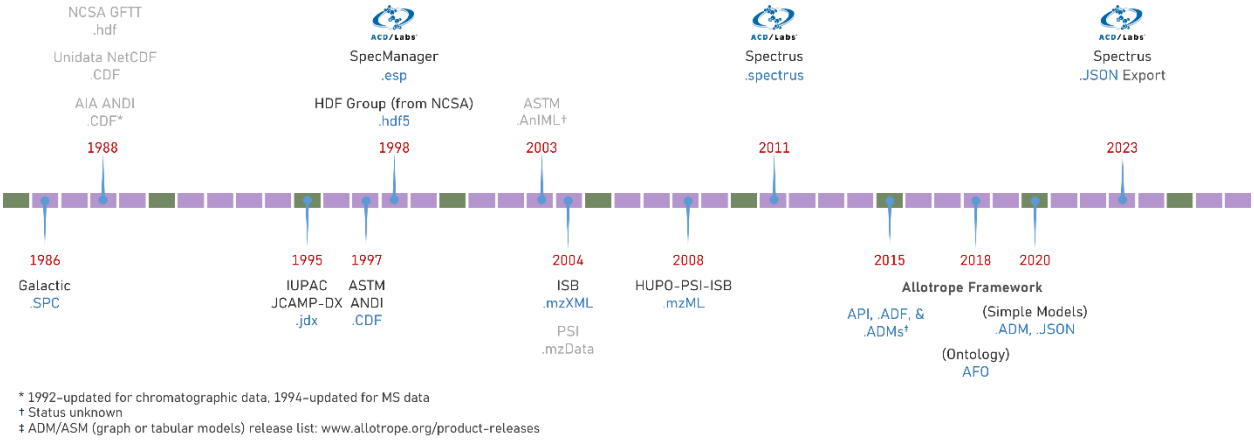

Figure 1. Notable analytical data standardization efforts over the decades (grey indicates formats that are no longer in use, or appear to have stalled)[§]

[§] This image is intended as an illustration of the major analytical data standardization efforts and is not an exhaustive summary.

Initial attempts, such as Galactic's SPC format released in 1986, were followed by formats like JCAMP-DX and ASTM's ANDI standard (mid-1990s). Increased focus on data digitalization, data management, and remote data access in the 2000s led to the formation of committees, industry consortia, and collaborative frameworks to help address analytical data standardization. The eXtensible Markup Language (XML) based AnIML (Analytical Information Markup Language) is one example proposed in 2003. More recently, the Allotrope Foundation—a consortium of organizations from the pharmaceutical industry—established an Analytical Data Ontology (ADO) and created their own data format in the form of Allotrope Simpe Models (.ASMs).

ACD/Labs has contributed to analytical data standardization and integration since 1998, first establishing the proprietary .esp format for the SpecManager platform. This was superseded by the .spectrus format with the launch of the Spectrus Platform in 2011, incorporating HD5 technology as part of the .spectrus specification. As with its predecessor, the Spectrus format supports a broad range of proprietary analytical instrument vendor formats. Natively supporting more than 150 formats that span from legacy instrument formats to the most up to date formats from major instrument vendors. Spectrus covers all major spectral and chromatographic data, widely used open standards, and is continually evolving to accommodate the changing landscape of formats and instrument innovations.

Open Data Formats vs. Proprietary Data Formats

"Non-proprietary or proprietary?" is a common discussion and data formats are no exception. Both have their advantages. Non-proprietary formats can be published under a wide range of licensing conditions. Being fully publicly disclosed often leads to being denoted an "Open Data Format" but there is no guarantee they are free (no cost) to use by everyone.

Open formats are usually supported by a standards body or community, and some offer free access and unrestricted usage. They enable data accessibility, longevity, and software interoperability often desired from standardized data. Selecting an open format could also represent cost savings (no need to purchase proprietary software) but you may need to make provisions to convert existing/legacy data or update automated workflows. Other considerations when selecting open formats for analytical data are to strategically choose formats to accommodate all your data, as no single open data format supports all analytical data types, and to understand what you may lose in meta data by converting data from a proprietary format to an open format.

Proprietary data formats may be considered high-risk. You are locked in with the vendor for access to your data through licensed products that represent continued expenditure and are at risk of obsolescence. Proprietary data formats also offer advantages, however. They enable analytical instrument vendors the freedom to innovate and deliver cutting edge technologies to R&D. Proprietary formats deliver rich, highly specific functionality and may provide some limited export of data from their formats.

How Open & Third-Party Formats Access Proprietary Vendor Data

Scientific instrument vendors often provide software development kits (SDK's) for their proprietary data format. Open formats and third-party proprietary data formats depend on these SDKs to translate vendor data into their own format, but data access can be limited. There can be delays in the update of Vendor SDKs made available to third parties and bugs in SDK's can result in errors translating data.

A Format for All Analytical Data

Spectrus is a proprietary analytical data format that bridges the gap between instrument vendor proprietary formats and accessible, standardized analytical data. Building on more than 25 years of experience, the Spectrus format is actively maintained by expert software developers and scientists working with spectroscopists, chromatographers, and analytical scientists at the cutting edge of R&D. Active partnerships between ACD/Labs and major instrument vendors ensure up to date format support and the capability to include new formats as they emerge. Spectrus also supports open formats and legacy formats (increasing the lifespan of instruments no longer supported by the vendor and ensuring value can be extracted from legacy data). Finally, conversion of historical data can be supported through engagement with our professional services team, or we can provide you with the tools for this undertaking.

Data Standardization for AI/ML

As R&D increasingly embraces AI/ML technologies, the need for robust data engineering practices has become paramount. Data standardization is a critical step in engineering data. AI/ML necessitates a focus on machine-readable, tokenizable, scalable, and interoperable formats.

JSON (JavaScript Object Notation) has emerged as a popular format for data ingestion in AI/ML workflows. JSON's rise as the preferred format for AI/ML workflows can be attributed to its flexibility and widespread compatibility. Its lightweight, human-readable structure is particularly well-suited for representing complex hierarchical data, such as chemical properties and analytical results. JSON's ability to encapsulate both data and metadata allows for unambiguous representation of information, making it a natural fit for workflows in chemical informatics and analytical sciences. Furthermore, its ease of integration with modern tools and APIs ensures compatibility with cloud-based AI/ML platforms. In addition, the ubiquity of JSON parsers in programming languages, such as Python and R, simplifies its adoption across diverse systems.

However, JSON is not without limitations. Its verbosity can pose challenges for large datasets, leading to inefficiencies in storage and transmission. This issue becomes particularly pronounced when dealing with high-volume analytical data, such as spectroscopic or chromatographic results. You may want to explore alternative formats that may be better suited to the unique demands of chemical and pharmaceutical AI/ML applications.

Domain-specific standards play an important role in the chemical and pharmaceutical industries. Extensible Markup Language (XML) based formats can be designed to support analytical instrumentation data, enabling high-fidelity data sharing across platforms. Similarly, Chemical Markup Language (CML) is tailored for chemical informatics, offering specialized features for

representing molecular structures, reactions, and spectra. These standards ensure compatibility with existing tools and workflows while addressing the unique needs of chemical R&D.

Balancing Analytical Data Standardization for the Lab and Data Science

The adoption of versatile data formats, such as JSON, highlights the increasing need for flexibility in AI/ML workflows. However, meeting the varied requirements of analytical chemistry necessitates a balance between general-purpose formats and domain-specific standards, ensuring high quality re-usable data.

Domain-specific formats are designed to address industry-specific challenges and should be leveraged whenever possible. Multi-format compatibility, as delivered by JSON formats, should be prioritized to accommodate AI/ML workflows. In addition, metadata management must be integrated into all stages of data engineering to ensure data provenance, reproducibility, and compliance with regulatory requirements. Finally, tools for automatic data validation should be implemented to prevent errors and streamline data ingestion into AI/ML pipelines.

Contextualization: Assembled Chemical & Analytical Data

The assembly of analytical data from multiple datasets is an essential practice in modern chemical studies and molecular characterization. Unlike a single dataset which typically provides information limited to a specific method or measurement, the assembly of multiple datasets enables a more comprehensive and multidimensional understanding of a chemical system. This approach is particularly valuable when addressing the complexity of real-world samples or intricate molecular structures, as each dataset contributes unique and complementary information.

In molecular characterization, datasets from techniques like NMR, LC/UV/MS, and IR spectroscopy can collectively describe structural, spectral, and compositional properties. Individually, each dataset might highlight one or more specific features, such as molecular connectivity, mass, or functional groups. When these datasets are combined, however, they provide a holistic picture that improves confidence in structure elucidation, facilitates the identification of unknown compounds, and enables the exploration or affirmation structure-property relationships.

Similarly, in degradation studies, combining data from LC/UV/MS, NMR, and other techniques enables the identification of degradation pathways, the characterization of degradation products, and the quantification of their rates of formation. By assembling these datasets, researchers can establish a comprehensive understanding of the chemical and environmental factors influencing stability and performance.

By assembling and analyzing such datasets collectively, researchers gain a more reliable and detailed understanding of the processes under investigation. This holistic view supports more informed decision-making. Moreover, integrated datasets provide a robust foundation for modeling, prediction, and AI/ML workflows, where the relationships between chemical behavior and external factors can be more effectively explored and utilized.

Analytical Data Standardization: The Foundation for Future Innovation

Data standardization is a pre-requisite for unlocking the full potential of analytical data. With its diverse techniques and proprietary instrument formats, analytical chemistry exemplifies the challenges and rewards of data standardization.

Data standardization harmonizes data, facilitates centralized management and access, and supports advanced analytics. Simply aggregating heterogeneous data in the cloud, data lakes, or data warehouses does not make it useable. R&D organization must create future-proof repositories that are amenable to scientists and help unlock the transformative potential of AI and ML. Selecting the right data standard means choosing between open or proprietary formats; the lightweight, human and machine-readable JSON format, or domain-specific formats. Assemblies of multi-technique datasets will help organizations gain a holistic view of chemical processes, enhance efficiency, regulatory compliance, and innovation.

Authors

Richard Lee is the Director of Core Technology and Capabilities at ACD/Labs. He obtained his Ph.D. from McMaster University focusing on strategies for metabolite identification and metabolomics studies, followed by research developing radiopharmaceuticals as imaging agents and therapeutics for oncology.

Sanji Bhal is the Director of Marketing & Communications at ACD/Labs. She completed her Ph.D. in synthetic organic chemistry at the University of Reading, a post-doctoral fellowship at Cancer Research UK, and worked as a medicinal chemist for several years