From silo to lake: Data standardization and the analytical laboratory

The pace of data collection has never been faster. "Data has just exploded in the laboratory environment over the years," says Clive Higgins, head of Connected Lab at MilliporeSigma. "Data files have gotten bigger and more complex." Automation and high-throughput techniques enable teams to collect more comprehensive data and metadata at speed, and advanced computational methods offer an exciting opportunity to extract insights from those data at scale.

But organizing and interpreting these rapidly growing data sets can be a major challenge. Most labs have numerous analytical instruments, each one recording a different type of data in its own format. Even instruments that collect the same measurements may store them in incompatible proprietary formats. For scientists trying to assemble data from multiple analyses, or later trying to discern themes in large sets of experimental records, the need to convert between file types to make comparisons or conduct analyses becomes an obstacle. "The challenge hasn't gone away, it's just gotten bigger," says Higgins. "The challenge has never been solved in any comprehensive way, until data standards came into place."

Software platforms that can integrate data from numerous sources and instruments and store it all in an organized, searchable form are increasingly important in modern laboratories. Many teams have developed data formats to standardize analytical data, working to streamline data collection, analysis, and storage.

THE CASE FOR STANDARDIZING DATA

A 2015 survey at GlaxoSmithKline found that researchers almost unanimously had difficulty working with data collected by other teams. That limitation contrasted starkly with the company's strategic goal of integrating its expanding capacity to collect genetic sequences, deep data from clinical trials, and real-world data from sources such as health records.[1]

While that survey focused on clinical data, the situation is no different in the laboratory. Siloed and fragmented data, produced on different instruments by members of different teams, impede a company's ability to store data long term and reuse it for future exploration. Moreover, when groups such as a contract research organization and the commissioning company collaborate, differences between their instruments or software products might make data transfer difficult.

Claudia Schwarz, associate director of informatics marketing at MilliporeSigma, says labs have increasingly realized that old data is "not just a bycatch of analytical measurement anymore" but becomes a real asset and source of information that their owners can learn from. But for pharmaceutical and chemical companies to build efficient and predictive analytical tools that take advantage of the many types of data they collect constantly, the data must first be comparable to one another. Harmonizing all sorts of data into a standardized file type that encodes them in an agreed-on way is the solution that many software development teams propose.

Proponents of standardized data formats argue that widespread adoption of standards will make the data management ecosystem easier for all parties: not only for laboratories commissioning or conducting analyses but also for instrument vendors and informatics software developers.

For example, if five software tools each needed to import data stored in any of five different instrument-specific formats, building a one-to-one converter for each pair would result in 25 data converters. On the other hand, if the hypothetical instrument manufacturers and software developers could agree on a common language, their systems could be integrated at much lower cost.

Such interoperability is one pillar of the FAIR data principles — findable, accessible, interoperable, and reusable — that serve as a guiding framework for managing scientific data to make them as useful as possible.[2] Using a standardized format can address the requirements for interoperability, which means that data tools from competing enterprises should be able to work together without too much user engineering; and for reusability, which requires that any metadata are packaged with the data and describe them thoroughly enough to make sense even in a new context.

Annette Hellbach, MilliporeSigma's head of product at Connected Lab, says standardization "enables you to take decisions in a faster and

more reliable manner, because for the first time you have all the data in front of you in one format."

Standardization can also streamline long-term storage in regulatory environments that require data to be kept for years. Maintaining many types of software over decades can be costly, says Higgins. If all the data are stored in one format, he says, then just one tool for reading that format has to be maintained. Another benefit of standardized data is that users can store and use that information independent from its source of origin, which means that even if a vendor discontinues an instrument or software product, the data are safe.

NUMEROUS DATA FORMATS

Technique-specific standardization

Many research communities use standard formats that were designed to handle a particular type of analytical data. For example, mzML is an open data format for mass spectra that the Human Proteome Organization's Proteome Standards Initiative created to make it easier for labs to share and compare experimental mass spectrometry data. The development team merged two earlier data formats to define metadata about an experiment, such as instrument descriptors and data-processing details, that should always be packaged with the raw mass spectrometry data. Its developers also eliminated redundant ways of encoding the same information.[3] Most mass spectrometers now support data export in mzML format.

Similar standardized formats exist for chromatography, nuclear magnetic resonance, diagnostics, medical imaging, and many other techniques. While such technique-specific standards solve the problem of incompatible proprietary formats, they cannot integrate data from different research approaches.

> "Not having data standards is not the problem. Instead, we have too many data standards with too little reach in the industry."
>
> Annette Hellbach,
> head of product at Connected Lab at MilliporeSigma

"When you're talking about characterizing materials, often you need more than one piece of equipment," Graham McGibbon, the director of strategic partnerships at the software company ACD/Labs, says. "Aggregation and data analysis and bringing it together then becomes the next bottleneck."

To address this bottleneck, numerous groups have developed generic formats to accommodate data from many types of analytical chemistry techniques, including chrmatography, mass spectrometry, NMR, and ultraviolet and infrared spectroscopy. Some of these formats use text-based markup languages, which a human can read, while others rely on faster-to-parse binary formats for better processing speed. Some are open source, others proprietary. "Not having data standards is not the problem," Hellbach says. "Instead, we have too many data standards with too little reach in the industry."

FORMATS MADE FOR MULTIPLE ANALYTICAL TECHNIQUES

Chemical JSON

Part of the reason so many data formats exist is that a number are built to solve specific, in-house problems. The Chemical JSON format is an open-source example that one team built to encode molecular structures. Marcus Hanwell, a staff scientist at Brookhaven National Laboratory who developed the standard, says, "I hesitated for years to make yet another standard. I kind of hobbled along with the ones I had." Eventually, however, practicality won out. At the time, Hanwell was working on Avogadro, a tool for visualizing and editing molecules that used JavaScript, C++, and Python languages for different functions. To speed up information flow between these modules, Hanwell and his colleagues developed a format based in the widely used JavaScript object notation, or JSON.[4]

A Chemical JavaScript object notation (JSON) file describes a molecule by listing its component atoms, their coordinates in 3D, and the bonds between them. It also contains metadata that can include chemical identifiers, along with physical properties such as boiling point or crystal unit structure; additional data layers can incorporate denser data like orbital structures.

While some chemical engineers and computational chemists have adopted Chemical JSON for quantum mechanical simulation or large-scale molecular docking, the intention was not to unite all computational chemists. "We're on the pragmatic side of formats," Hanwell says. "We had a niche we were working in [and] if other people found it useful, they could use it." That is a contrast to data formats developed by groups explicitly aiming to standardize their fields.

AnIML

Analytical Information Markup Language (AnIML), is an open-source format that was developed by a community of scientist and noncommercial stakeholders to harmonize multiple analytical data formats. It was commissioned in 2003 by the American Society for Testing and Materials (ASTM), which wanted a standard data format for analytical chemistry and biology.[5] The working group used earlier efforts to standardize analytical data as a starting point to develop AnIML in the text-based programming language XML, which is readable by both humans and machines.

The development team built a flexible format by setting up a two-layer system. First comes a data definition, which defines what attributes a certain type of file should have; second is a data package that contains the actual data and metadata, organized as the data definition suggests. When new techniques are invented, AnIML can simply develop new data definitions, according to Burkhard Schaefer, a leader in the format's continuing development. "We didn't want to build a data format that only works for the set of use cases we had on the table."

Schaefer says the committee's approach for each new technique is "to say to subject matter experts, 'What does your data look like?' And then we listen." Following ASTM convention, the AnIML working group requires consensus from stakeholders before each new data definition is approved. According to Schaefer, the process takes a little time but produces robust results. AnIML currently supports data from many analytical chemistry data types, such as those from chromatography and mass spectrometry, and is expanding into bioprocessing and biochemical techniques such as PCR.

ADF

The Allotrope Data Format (ADF) was designed with pharmaceutical companies' analytical data in mind. A coalition of about a dozen companies launched the Allotrope Foundation in 2012 "to make the intelligent analytical laboratory a reality."[6] Wolfgang Colsman, a former lead architect for Allotrope, says that the design brief for the format was—in contrast to technique-focused standard formats—to "provide a solution for everything."

The format depends on a comprehensive vocabulary of terms describing an experiment or sample's attributes. For each analytical method the ADF supports, a data model indicates which attributes to expect. The values of those attributes are stored in a data package, along with ancillary files and metadata. The format uses a binary programming language to enable fast computation.

The ADF currently supports data from various techniques for analytical chemistry and process monitoring, and framework development is supported through annual membership and license fees.

USING STANDARDIZED DATA

Accessible, interoperable data become available for reuse in ways that might not have been predicted when it was first collected. Arne Kusserow, a product manager at MilliporeSigma, says, "It's nice to have things standardized, but by itself it's worth nothing. But when you have all your data in one place and all the data is standardized, you can apply other techniques," such as searching and visualization, or machine learning and other advanced techniques for exploratory analysis.

Many companies are developing software platforms to analyze and manage standardized data from multiple techniques. ACD/Labs sells a data management platform that relies on a data format of its own to integrate chemical structures with diverse analytical data in one interface.[7] MilliporeSigma also offers a knowledge management platform, called BSSN Software, which is based in AnIML[8]. Hellbach says it includes data visualization and long-term archiving functions, along with a platform and AnIML converters that enable customers to feed data into their data lake in an organized and standardized way.

A data lake—a cloud-based, company-wide data repository—is the strategic endpoint of data standardization. Getting there can be a formidable undertaking, because of the volume of data that must be organized and the multitude of formats. At GlaxoSmithKline, a data lake comprising clinical trial data, genetic sequences, and electronic health records took months to populate and involved half a dozen external vendors.[1] Partly because of such challenges, experts say the industry is still early in the adoption curve, although interest in data standardization has grown.

Companies that are comprehensively digitized have a strategic advantage over competitors that have not yet transitioned or have begun in a piecemeal way, according to Hellbach. "In the end," she says, "data standardization helps the whole value chain, starting on basic levels of the lab up to business leadership."

> "It's nice to have things standardized, but by itself it's worth nothing. But when you have all your data in one place and all the data is standardized, you can apply other techniques."
>
> Arne Kusserow, product manager at MilliporeSigma

References

1. Thomas H. Davenport and Randy Bean, "Biting the Data Management Bullet at GlaxoSmithKline," *Forbes*, Jan. 8, 2018, https://www.forbes.com/sites/tomdavenport/2018/01/08/biting-the-data-management-bullet-at-glaxosmithkline/?sh=1f10f6325577.

2. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Sci. Data* 3 (March 15, 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

3. Erik W. Deutsch, "Mass Spectrometer Output File Format mzML," *Methods Mol. Biol.* 604 (2010): 319–31, https://doi.org/10.1007/978-1-60761-444-9_22.

4. Marcus Hanwell et al., "Open Chemistry: RESTful Web APIs, JSON, NWChem and the Modern Web Application," *J. Cheminf.* 9 (Oct. 30, 2017): 55, https://doi.org/10.1186/s13321-017-0241-z.

5. Burkhard A. Schäfer et al., "Documenting Laboratory Workflows Using the Analytical Information Markup Language," JALA 9, no. 6 (Dec. 1, 2004): 375–81, https://doi.org/10.1016/j.jala.2004.10.003.

6. Allotrope Foundation website, "About Us," https://www.allotrope.org/about-us.

7. Spectrus white paper, "Looking Beyond Analytical Data Standardization—the Fourth Paradigm," ACD Labs, https://www.acdlabs.com/comm/data_standardization/whitepaper/.

8. AnIML white paper, "The Need for a New Data Standard," MilliporeSigma, https://bssn-software.com/capabilities/animl-data-management/.

DATA MANAGEMENT SOLUTIONS

unleash the power of laboratory data

For better data integrity, data management and scientific collaboration

Data Management Solutions from MilliporeSigma (formerly BSSN™ Software) lets you access, integrate and share analytical and biological lab data in a structured format for greater lab standardization and interoperability. Its vendor-neutral platform helps you overcome accessibility limitations in lab data storage by eliminating traditional paper-based or partially digitalized approaches.

With Data Management Solutions, you can easily integrate your instruments with leading laboratory software systems like LIMS or ELN and facilitate data exchange with internal or external partners.

- Convert proprietary instrument data into an open standard
- Remain instrument vendor agnostic
- Enable flexibility and straightforward data integration
- Practice FAIR (Findable, Accessible, Interoperable, Reusable) data principles

AnIML

AnIML is the open source ASTM XML standard for analytical chemistry and biological data. An AnIML file consists of the following:

- A Core Schema that describes a set of rules defining the valid structure of an AnIML file.
- A Technique Schema that describes how to handle and render domain-specific data in a way that makes sense to the user for a given technique.
- A set of Technique Definition documents that tightly constrain the Core Schema and are defined by the Technique Schema. You can extend Technique Definitions to accommodate vendor- and institution-specific data fields.

Many tools for XML are available off the shelf, making AnIML implementation easier. Because XML is text based, AnIML documents are human readable, which is critical to long-term storage.

data integrity

Capture & Convert

- AnIML converters for 300+ instruments and data formats
- Bringing analytical and biological data into one common format

Integrate

- Provide instrument data to LIMS, ELN, LES or SAP for consumption

data management

Collect

- Data lake as home to all laboratory data and to all metadata-driven navigation & queries
- Storage (cloud or on-premise) for analytical and biological source data, processed data and metadata

Visualize & Process

- Technology-neutral visualization and processing of converted data and metadata
- Transparently review your data any place, any time
- Windows, Mac, HTML

collaboration

Collaborate

- Collaboration with internal & external partners (CRO, CDMO)
- Get organized and maintain the same data quality as in-house

Consume & Analyze

- Reuse laboratory data for predictive analytics, statistics, data mining, and AI & ML
- Enable data intelligence and secondary use of data

archive

Retain Data

- Long-term retention of GxP-relevant data
- Storage of data over a long period of time in an accessible and human-readable format

applications

Analytical Chemistry

- HPLC
- IC
- CE
- GC
- UV/Vis
- Infrared, FTIR
- Raman
- Mass spectrometry
  - HPLC-MS
  - GC-MS
  - ICP-MS
- NMR
- XRD
- Laser diffraction

Biology

- Bioreactors
- Microbioreactors
- Microplate readers
- Cell counters
- Bioanalyzers
- Purification chromatography
- qPCR

Other

- Balances
- pH meters
- Imaging
- Pipetting/liquid handling
- Densiometry

Visit BSSN-Software.com