BSSN Software is now part of MERCK

AnIML: Concepts & Applications

The Analytical Information Markup Language

Burkhard Schaefer, Head of Core Technologies and Partnering,
Merck KGaA, Darmstadt, Germany

Synopsis

This whitepaper presents an overview of AnIML, the Analytical Information Markup Language. It describes its features and highlights a number of possible applications. It is intended to assist decision makers in the user and vendor community when evaluating the use of AnIML as a data format.

About the Author

Burkhard Schaefer is the Head of Core Technologies and Partnering at Merck KGaA, Darmstadt, Germany, and was the founder of BSSN Software.

He got involved in the AnIML project in 2003 when he was working in the Analytical Chemistry Division of the Chemical Science and Technology Laboratory at the National Institute of Standards and Technology (NIST). He is responsible for the design of the AnIML architecture as it stands today.

The life science business of Merck operates as
MilliporeSigma in the U.S. and Canada.

Welcome to AnIML

Meet AnIML – the Analytical Information Markup Language. AnIML is a standardized data format that allows for storing and sharing of experiment data in a single format. It is suitable for a wide range of analytical measurement techniques. Using AnIML, you can accurately record and document laboratory workflows and results, no matter which instruments or measurement techniques were used.

To achieve this, AnIML provides a generic data container – the **AnIML Core** – that permits the storage of arbitrary analytical data. This includes

- sample information,
- method information,
- measurement results,
- instruments and software used,

as well as workflow information that ties experiments and samples together. The concept of Technique Definitions permits the formal specification of constraints for using this data container. Such a definition can prescribe how the data for specific measurement techniques should be captured in an AnIML document. This way, AnIML can be applied to many different analytical techniques. It grows gracefully as new techniques are developed, allowing continued use of existing software components.

AnIML is based on XML. This has two interesting consequences. First, many tools for XML manipulation are readily available off-the-shelf, making implementation easier. Second, as XML is a text-based format, AnIML documents are human-readable – an important aspect for long-term storage.

While AnIML has its roots in analytical chemistry, efforts have been made to make the standard applicable to many other scientific domains. AnIML was developed by the ASTM E13.15 subcommittee on analytical data which consists of volunteers from industrial, academic, government and vendor communities.

The Need for a New Data Standard

One of the challenges in laboratory data management is the handling of experiment data. There are excellent instruments available from many different vendors, but most instruments produce data in their own proprietary formats. This leads to major difficulties for data processing, sharing, and archiving. There is little choice in software, as users are often tied to the tools that came with the instrument.

Long-term storage

By its very nature, every electronic record carries an implicit expiration date. The reason for this is that the system components required for data access may degrade or become unavailable over time. Archiving data in a proprietary format means that the original application is required to open the data in the future. Accordingly, organizations need to retain this software for the desired lifetime of the data. As hardware and operating systems evolve, such software will no longer run. Access to the data is lost. Users are forced to buy software upgrades for every type of instrument data they need to maintain. This is costly, and only works as long as the chosen data formats are still supported – and the vendors still is in business. Through the use of techniques like virtualization and hardware emulation, lifetime of software can be extended, reducing the frequency of upgrades. However, this only defers the problem.

The solution is to move the data into an open, standardized, multi-technique format. This greatly reduces the number of software tools to maintain to retain access to the data.

Data interchange

Due to the proprietary nature of instrument data files, it is difficult to share data with other scientists. A number of standards already exist to address this issue. These include SpectroML, JCAMP-DX and ANDI. However, these standards only apply to particular measurement methods. They are not intended for cross-technique applications.

This makes data sharing, analysis, collaboration, instrument integration, and interfacing with other software such as laboratory information management systems (LIMS) difficult.

Having a universally-accepted standard that supports multiple techniques makes data exchange much easier.

Experiment complexity

Experiments in an automated laboratory are becoming more and more complex, involving the use of multiple analytical techniques. These techniques are often applied in combination or in sequence on the same sample. Workflows may also involve sample acquisition, sub-sampling, and sample preparation. At the moment, no other file format exists which can represent such experiments. Before AnIML, this meant that a home-grown custom data representation needed to be devised and implemented to support such experiments.

A Look at the Architecture

We‘ll now explore the architecture behind AnIML at a high level. The standard is divided into two logical layers, the **AnIML Core** and the **AnIML Technique Definitions**. Each of these layers is described by an XML Schema, the **Core Schema** and the **Technique Schema**.

AnIML core

The AnIML Core provides a universal container for arbitrary analytical data. This container is very flexible. It accepts name-value pairs, hierarchies and multi-dimensional data sets to represent the actual data. Additionally, mechanisms for organizing the data into samples and experiments are provided. Every AnIML document is governed by the Core Schema.

AnIML technique definitions

To allow for interchangeability of the data, it is necessary to constrain the usage of the Core. That's where Technique Definitions come in. A Technique Definition describes how to use the Core for experiments of a particular measurement technique. Like a digital blueprint, it defines how the data need to be structured and labeled. This description is machine-readable: Technique Definitions are simple XML documents, governed by the Technique Schema.

AnIML technique extensions

Sometimes, the fields defined in a Technique Definition are not sufficient to characterize an experiment. This can be the case if an instrument can measure additional parameters, or when users need to store additional sample information. To accommodate such additional fields, a **Technique Extension** can be used. A Technique Extension defines which fields should be added to a technique and how they are structured. As this description is a machine-readable XML document, software can discover such additional fields programmatically. This way, the standard can be extended without breaking compatibility with existing applications.

Inside the Universal Data Container

The AnIML Core can be considered the heart of the standard. With its flexibility, it provides a universal data container which is used to document our experiments. This section will highlight some of the concepts of the Core. Each AnIML document can contain **Samples**, **Experiment Steps**, an **Audit Trail**, and **Digital Signatures**.

Samples

Each sample used in an AnIML document is declared at the top of the document. It is identified by a sample ID, a bar code, and a name. Samples can carry any number of attributes. These attributes can be organized in hierarchical categories, allowing for complete representation of sample characteristics.

The attributes which describe a sample vary depending on the experiment performed. To be precise: They are specific to the analytical technique. A Technique Definition can therefore prescribe the structure of a sample definition.

If required, the relationship between a sample and its container can be described as well. This is useful if your samples are located in microtiter plates, racks, vials, or on 2D gels. Each sample is declared only once per document. It can then be referenced using its sample ID.

Experiment steps

An Experiment Step documents how a particular analytical technique has been applied to a technique-specific set of samples. It can be thought of as an instance of a Technique Definition, or as the basic building block in an analytical workflow. An Experiment Step always represents a single action, such as the acquisition of a single infrared spectrum.

For each Experiment Step, the AnIML Core can store the measured results, a description of the method, references to the samples involved, and the Technique used. It also provides a number of metadata fields that further describe how the experiment was performed. The exact content is technique-specific and prescribed by the corresponding Technique Definition.

Experiment Steps can both consume and produce samples. Analytical techniques typically consume samples and analyze them. However, sample preparation or separation techniques also produce new materials. An example would be liquid chromatography coupled with a mass spectroscopy (MS) detector. A chromatography Experiment Step consumes the sample, separates it, and produces fractions. Each time a mass spectrum is acquired, a corresponding MS Experiment Step is created which consumes and analyzes the fraction as it exits the column. With this mechanism, material flow between experiments can be tracked.

An Experiment Step also can consume data produced by other Experiment Steps. Let's consider an MS Experiment Step with a mass spectrum result in it. This spectrum is processed and generates a peak table. Accordingly, a new peak table Experiment Step is created which consumes the data measured by the MS Step. This illustrates the data flow between experiments.

Representing Laboratory Workflows

AnIML expresses laboratory workflows by assembling and connecting Samples and Experiment Steps. This way, both the material and the data flow in the process are captured. This may sound complex. However, a document only gets as complex as the workflow it represents. Simple experiments are straightforward to represent using AnIML.

A simple infrared experiment

Analyzing a sample with an infrared spectrophotometer generates two entries in an AnIML document. The first one is a sample definition which contains a unique sample ID and the sample attributes. Additionally, a single Experiment Step is created. It contains the infrared spectrum, the instrument parameters, the sample ID from above, and a reference to the Infrared Technique Definition.

Enriching the workflow with other techniques

Having started with a simple experiment, we can now expand the workflow. Let's assume that the same sample is now analyzed using UV/Vis spectroscopy. A new Experiment Step is then created, containing the UV/Vis trace. It is configured to reference the same sample ID as the infrared Step above.

If desired, additional Experiment Steps could be added to describe the sample preparation.

Making AnIML Work With Any Measurement Technique

The fact that the AnIML Core is independent of measurement techniques makes AnIML very powerful. With such a generic architecture, AnIML can handle data from all well- known and frequently-used techniques, including spectroscopy, chromatography, imaging, and others. However, it is also possible to use it for custom or one-off experiments, micro-fluidic chips or special sensors, making these techniques first-class citizens in the data system. Over time, new analytical techniques and their corresponding Technique Definitions will be developed. This generic approach allows the system to use them without requiring modifications or software upgrades.

As discussed above, Technique Definitions and Technique Extensions provide a machine-readable definition of the layout of Experiment Steps and Samples. Not only does this allow software to discover how to create data files for a given technique. It also permits validating the content of an AnIML document. Here, the Technique Definitions serve as a check list of data fields that need to be present. Technique Definitions and Extensions can mark these fields as optional or as required. This way, a validator tool can verify that all required information about an experiment has been captured properly. This opens new possibilities for improved data quality and integrity.

Regulatory Compliance

Laboratories involved in certain aspects of pharmaceutical or environmental research are subject to certain rules and regulations, such as 21 CFR Part 11, EPA CROMERR, Good Laboratory Practices (GLP), and Good Manufacturing Practices (GMP). AnIML provides mechanisms to help organizations address these requirements.

Digital signatures

Certain regulations require the implementation of electronic signatures on data records. AnIML supports this by providing a standardized way of applying digital signatures to scientific data. Like all data formats, AnIML is not in a position to prevent unauthorized modification of a document. This is the job of the underlying storage system, where appropriate access controls need to be enforced. However, using digital signatures, such unauthorized changes can be detected.

Besides detecting modifications, it is also possible to use signatures to implement sign-off workflows and multi-level approval processes in a paperless laboratory.

Users can apply multiple signatures to an AnIML document. Each signature can have a different scope, covering different parts of a document. This proves quite useful in practice: Each staff member can sign for those parts of the experiment he or she conducted. Eventually, one could envision instruments directly signing the results they produce, proving data authenticity. Such a mechanism allows for total traceability of scientific data – from the long-term archive all the way back to the original instrument.

The AnIML signature mechanism leverages the established W3C XML Digital Signature standard. This ensures compatibility with existing public key infrastructures, certificates, key tokens, and smart cards. In many countries, it is possible to create legally binding signatures this way.

Recording changes

Changes to AnIML documents can be recorded in the built-in audit trail. Each audit trail entry accurately records all aspects of a change. This includes the old and the new values, the person responsible, the reason for the change, as well as a time stamp. Using the audit trail, it is possible to examine the changes and revert the document into previous stages. To increase security, audit trail entries can be digitally signed.

Application Areas

We will now explore a number or application areas where the deployment of AnIML promises significant advantages.

Long-term archiving

Depending on the use case, electronic data records need to be preserved for several decades. Today, many laboratories choose to archive their data in PDF format. However, PDF only captures an image representation of the scientific data, not the actual raw numbers. This makes it difficult to perform certain operations (re-processing, re-calculations, re-integration) on such files. AnIML ensures that the actual values are preserved and remain readable and available for reprocessing in the future.

It is no longer necessary to retain the original instrument software. Instead, a single generic AnIML viewer tool can be used to access any document in the archive. This results in significant cost savings when maintaining the archive.

At its core, AnIML stores the experiment data in XML. This ensures that the data is structured and tightly constrained, but human-readable at the same time. Thus even if the software to access the archived data is lost, full reconstruction of any archived record remains possible. Also, the syntactic (XML) and semantic (technique) integrity of archived records can be verified with a validation tool to maintain data quality and archive integrity.

Cross-technique data mining

Having data of multiple techniques in the same format allows cross-technique data analysis and provides a solid foundation for data mining tools.

Workflow standardization

Instrument software today offers many features for analysis and post-processing of result data. However, every vendor implements algorithms slightly differently. This makes it hard to compare results from different instruments. To solve this problem, organizations can convert results into AnIML as soon as they are produced by the instrument. Any processing and analysis can then be performed on the AnIML document using a standard tool, improving consistency.

Data publication

AnIML can be used to publish scientific data on the Internet or Intranet. Journal publishers and authors can make supporting materials for articles available for download in AnIML format. Some publicly funded research programs require submission or public dissemination of results. AnIML can serve as a valuable tool for such efforts.

Software consolidation

By design, AnIML allows the creation of generic software components that work with any analytical technique. Such software is independent of the measurement technique, the instruments used, and the software needed to drive the instruments.

Accordingly, we only need to deploy and maintain a very compact set of software tools to the user's PCs. It is no longer necessary to install proprietary and expensive instrument software on all laboratory PCs to be able to simply view the data. The generic components are all that is required.

This aspect is also interesting for instrument vendors. They can reuse the same set of software components across their product lines, resulting in significantly reduced implementation time.

LIMS integration

Organizations looking to store certain parts of the result data in a laboratory information management system (LIMS) can also benefit from AnIML. Rather than implementing a LIMS interface for every instrument used, a single AnIML interface could be established. The LIMS would then extract any required information from the AnIML document, no matter which instrument was used.

Result delivery

With AnIML, laboratories have an open format at their disposal to deliver experiment results to end users. This is especially interesting for service labs which perform analyses on behalf of outside customers or other departments. The ability to deliver raw data in addition to PDF reports increases the confidence in results and allows re-processing by the recipient.

BSSN Software and AnIML

BSSN Software from Merck has developed a comprehensive offering of products and services to help organizations of all sizes introduce AnIML into their processes. In addition, we support instrument manufacturers in implementing AnIML in their products.

Due to our constructive involvement and technical leadership in the AnIML development process, BSSN Software is uniquely positioned to deliver AnIML-based solutions very efficiently. Today.

BSSN Workbench

View, manipulate and validate AnIML documents, technique definitions and extensions.

- Provides a 360° view on a sample
- Graphical workflow representation
- Dynamic Document Streaming for large files
- Flexible cross-technique reporting
- Easy LIMS and ELN integration

bssn-software.com/workbench

BSSN Hub

Efficiently store and organize all your laboratory data for quick and easy retrieval.

- Sophisticated metadata management
- Dynamic document streaming for large data sets
- Quick and easy retrieval
- Designed with long-term archiving in mind
- Easy integration with open and well-documented interfaces
- Works stand-alone or together with your existing scientific data management system (SDMS)

bssn-software.com/hub

SciDiver

Dive into your laboratory data with anyone, anywhere, on any device. SciDiver is a low-touch, self-serve solution for connecting your instrument data and collaborating with colleagues in the cloud.

bssn-software.com/scidiver

AnIML converters

Easily convert your proprietary instrument data into AnIML. Converters are available for many popular instruments and standard data formats.

bssn-software.com/converters

Services

Leverage our expertise for your next AnIML project. The following services are available.

- Data management strategy development
- AnIML implementation and roll-out
- Legacy data migration
- Custom converter development
- Validation of AnIML implementations
- Technique Definition / Extension development
- LIMS and ELN integration
- Training and coaching
- Consulting

bssn-software.com/services

OEM offerings

We realize that many instrument and software vendors are looking for strategies for adopting AnIML. BSSN Software's professional services can help with planning and implementation. Most of our tools are available for OEM licensing and have various branding options. Additionally, we offer software components that make AnIML implementation painless.

bssn-software.com/oem

Get started!

Interested in learning more about AnIML? Contact us today for a consultation. The lab informatics specialists from BSSN Software offer remote or on-site workshops designed to familiarize you with the details of the AnIML standard and help you to evaluate it. Together, we examine where it makes sense to integrated AnIML into your laboratory processes. After this workshop, you will know how AnIML works and what it takes to adopt it in your organization.

bssn-software.com

Top Ten Reasons for Using AnIML

1. **Open.** Independent of individual vendors.
2. **Technique-agnostic.** A single format captures data from any measurement technique.
3. **Extensible.** User-, vendor-, and instrument-specific content can easily be accommodated without negatively impacting compatibility.
4. **Engineered with long-term archiving in mind.** Data is human-readable. No more proprietary software needs to be maintained.
5. **Complex experiments.** Workflows of arbitrary complexity can be captured accurately.
6. **Regulatory compliance.** Audit trails, digital signatures and validation are available.
7. **Easy data exchange.** Get the right data to the right people at the right time.
8. **Cross-technique data analysis.** Analyze your data in new ways.
9. **Less software to deploy.** One single tool can handle arbitrary experiments.
10. **Great tools and services.** BSSN Software provides everything you need to deploy AnIML today.

Learn more at
BSSN-software.com

34374
01/2021

BSSN Software is now part of MERCK